Zbornik 17. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2014

Zvezek G

Proceedings of the 17th International Multiconference

INFORMATION SOCIETY – IS 2014

Volume G

Jezikovne tehnologije
Language Technologies

Uredila / Edited by
Tomaž Erjavec, Jerneja Žganec Gros

http://is.ijs.si

9.–10. oktober 2014 / October 9th–10th, 2014
Ljubljana, Slovenia

Zbornik 17. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2014

Zvezek G

Proceedings of the 17th International Multiconference

INFORMATION SOCIETY – IS 2014

Volume G

Jezikovne tehnologije

Language Technologies

Uredila / Edited by

Tomaž Erjavec, Jerneja Žganec Gros

http://is.ijs.si

9. – 10. oktober 2014 / October 9th - 10th, 2014
Ljubljana, Slovenia

Urednika:

Tomaž Erjavec
Odsek za tehnologije znanja
Institut »Jožef Stefan«, Ljubljana

Jerneja Žganec Gros
Alpineon, d.o.o.

Založnik: Institut »Jožef Stefan«, Ljubljana
Priprava zbornika: Mitja Lasič, Vesna Lasič, Lana Zemljak
Oblikovanje naslovnice: Vesna Lasič, Mitja Lasič

Ljubljana, oktober 2014

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

004.934(082)(0.034.2)
81'25:004.6(082)(0.034.2)

MEDNARODNA multikonferenca Informacijska družba (17 ; 2014 ; Ljubljana)
Jezikovne tehnologije [Elektronski vir] : zbornik 17. mednarodne multikonference Informacijska družba - IS 2014, 9.-10. oktober 2014, [Ljubljana, Slovenia] : zvezek G = Language technologies : proceedings of the 17th International Multiconference Information Society - IS 2014, October 9th - 10th, 2014, Ljubljana, Slovenia : volume G / uredila, edited by Tomaž Erjavec, Jerneja Žganec Gros. - El. knjiga. - Ljubljana : Institut Jožef Stefan, 2014

Način dostopa (URL): http://library.ijs.si/Stacks/Proceedings/InformationSociety

ISBN 978-961-264-077-4 (pdf)
1. Gl. stv. nasl. 2. Vzp. stv. nasl. 3. Dodat. nasl. 4. Erjavec, Tomaž, 1960-
275927552

PREDGOVOR MULTIKONFERENCI INFORMACIJSKA DRUŽBA 2014

Multikonferenca Informacijska družba (http://is.ijs.si) s sedemnajsto zaporedno prireditvijo postaja tradicionalna kvalitetna srednjeevropska konferenca na področju informacijske družbe, računalništva in informatike. Informacijska družba, znanje in umetna inteligenca se razvijajo čedalje hitreje. Čedalje več pokazateljev kaže, da prehajamo v naslednje civilizacijsko obdobje. Npr. v nekaterih državah je dovoljena samostojna vožnja inteligentnih avtomobilov, na trgu pa je moč dobiti kar nekaj pogosto prodajanih tipov avtomobilov z avtonomnimi funkcijami kot »lane assist«. Hkrati pa so konflikti sodobne družbe čedalje bolj nerazumljivi.

Letos smo v multikonferenco povezali dvanajst odličnih neodvisnih konferenc in delavnic. Predstavljenih bo okoli 200 referatov, prireditev bodo spremljale okrogle mize, razprave ter posebni dogodki kot svečana podelitev nagrad. Referati so objavljeni v zbornikih multikonference, izbrani prispevki bodo izšli tudi v posebnih številkah dveh znanstvenih revij, od katerih je ena Informatica, ki se ponaša s 37-letno tradicijo odlične evropske znanstvene revije.

Multikonferenco Informacijska družba 2014 sestavljajo naslednje samostojne konference:

- Inteligentni sistemi
- Izkopavanje znanja in podatkovna skladišča
- Sodelovanje, programska oprema in storitve v informacijski družbi
- Soočanje z demografskimi izzivi
- Vzgoja in izobraževanje v informacijski družbi
- Kognitivna znanost
- Robotika
- Jezikovne tehnologije
- Interakcija človek-računalnik v informacijski družbi
- Prva študentska konferenca s področja računalništva
- Okolijska ergonomija in fiziologija
- Delavnica Chiron.

Soorganizatorji in podporniki konference so različne raziskovalne in pedagoške institucije in združenja, med njimi tudi ACM Slovenija, SLAIS in IAS. V imenu organizatorjev konference se želimo posebej zahvaliti udeležencem za njihove dragocene prispevke in priložnost, da z nami delijo svoje izkušnje o informacijski družbi. Zahvaljujemo se tudi recenzentom za njihovo pomoč pri recenziranju.

V 2014 bomo drugič podelili nagrado za življenjske dosežke v čast Donalda Michija in Alana Turinga. Nagrado Michie-Turing za izjemen življenjski prispevek k razvoju in promociji informacijske družbe je prejel prof. dr. Janez Grad. Priznanje za dosežek leta je pripadlo dr. Janezu Demšarju. V letu 2014 četrtič podeljujemo nagrado »informacijska limona« in »informacijska jagoda« za najbolj (ne)uspešne poteze v zvezi z informacijsko družbo. Limono je dobila nerodna izvedba piškotkov, jagodo pa Google Street view, ker je končno posnel Slovenijo. Čestitke nagrajencem!

Niko Zimic, predsednik programskega odbora
Matjaž Gams, predsednik organizacijskega odbora

FOREWORD - INFORMATION SOCIETY 2014

The Information Society Multiconference (http://is.ijs.si) has become one of the traditional leading conferences in Central Europe devoted to information society. In its 17th year, we deliver a broad range of topics in the open academic environment fostering new ideas which makes our event unique among similar conferences, promoting key visions in interactive, innovative ways. As knowledge progresses even faster, it seems that we are indeed approaching a new civilization era. For example, several countries allow autonomous card driving, and several car models enable autonomous functions such as "lane assist". At the same time, however, it is hard to understand growing conflicts in the human civilization.

The Multiconference is running in parallel sessions with 200 presentations of scientific papers, presented in twelve independent events. The papers are published in the Web conference proceedings, and a selection of them in special issues of two journals. One of them is Informatica with its 37 years of tradition in excellent research publications.

The Information Society 2014 Multiconference consists of the following conferences and workshops:

- Intelligent Systems
- Cognitive Science
- Data Mining and Data Warehouses
- Collaboration, Software and Services in Information Society
- Demographic Challenges
- Robotics
- Language Technologies
- Human-Computer Interaction in Information Society
- Education in Information Society
- 1st Student Computer Science Research Conference
- Environmental Ergonomics and Psysiology
- Chiron Workshop.

The Multiconference is co-organized and supported by several major research institutions and societies, among them ACM Slovenia, SLAIS and IAS.

In 2014, the award for life-long outstanding contributions was delivered in memory of Donald Michie and Alan Turing for a second consecutive year. The Programme and Organizing Committees decided to award the Prof. Dr. Janez Grad with the Michie-Turing Award. In addition, a reward for current achievements was pronounced to Prof. Dr. Janez Demšar. The information strawberry is pronounced to Google street view for incorporating Slovenia, while the information lemon goes to cookies for awkward introduction. Congratulations!

On behalf of the conference organizers we would like to thank all participants for their valuable contribution and their interest in this event, and particularly the reviewers for their thorough reviews.

Niko Zimic, Programme Committee Chair
Matjaž Gams, Organizing Committee Chair

KONFERENČNI ODBORI
CONFERENCE COMMITTEES

International Programme Committee

Vladimir Bajic, South Africa
Heiner Benking, Germany
Se Woo Cheon, Korea
Howie Firth, UK
Olga S. Fomichova, Russia
Vladimir A. Fomichov, Russia
Vesna Hljuz Dobric, Croatia
Alfred Inselberg, Izrael
Jay Liebowitz, USA
Huan Liu, Singapore
Henz Martin, Germany
Marcin Paprzycki, USA
Karl Pribram, USA
Claude Sammut, Australia
Jiri Wiedermann, Czech Republic
Xindong Wu, USA
Yiming Ye, USA
Ning Zhong, USA
Wray Buntine, Finland
Bezalel Gavish, USA
Gal A. Kaminka, Israel
Mike Bain, Australia
Michela Milano, Italy
Derong Liu, Chicago, USA
Toby Walsh, Australia

Organizing Committee

Matjaž Gams, chair
Mitja Luštrek
Lana Zemljak
Vesna Koricki-Špetič
Mitja Lasič
Robert Blatnik
Mario Konecki
Vedrana Vidulin

Programme Committee

Nikolaj Zimic, chair
Franc Solina, co-chair
Viljan Mahnič, co-chair
Cene Bavec, co-chair
Tomaž Kalin, co-chair
Jozsef Györkös, co-chair
Tadej Bajd
Jaroslav Berce
Mojca Bernik
Marko Bohanec
Ivan Bratko
Andrej Brodnik
Dušan Caf
Saša Divjak
Tomaž Erjavec
Bogdan Filipič
Andrej Gams
Matjaž Gams
Marko Grobelnik
Nikola Guid
Marjan Heričko
Borka Jerman Blažič Džonova
Gorazd Kandus
Urban Kordeš
Marjan Krisper
Andrej Kuščer
Jadran Lenarčič
Borut Likar
Janez Malačič
Olga Markič
Dunja Mladenič
Franc Novak
Vladislav Rajkovič
Grega Repovš
Ivan Rozman
Niko Schlamberger
Stanko Strmčnik
Jurij Šilc
Jurij Tasič
Denis Trček
Andrej Ule
Tanja Urbančič
Boštjan Vilfan
Baldomir Zajc
Blaž Zupan
Boris Žemva
Leon Žlajpah
Igor Mekjavić
Tadej Debevec

KAZALO / TABLE OF CONTENTS

Zbornik 17. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2014

Zvezek G

Proceedings of the 17th International Multiconference

INFORMATION SOCIETY – IS 2014

Volume G

Jezikovne tehnologije

Language Technologies

Uredila / Edited by

Tomaž Erjavec, Jerneja Žganec Gros

http://is.ijs.si

9. – 10. oktober 2014 / October 9th - 10th, 2014
Ljubljana, Slovenia

PREDGOVOR K ZBORNIKU DEVETE KONFERENCE "JEZIKOVNE TEHNOLOGIJE"

V pričujočem zborniku so objavljeni prispevki z IS-JT 2014, devete konference "Jezikovne tehnologije", ki je potekala 9. in 10. oktobra 2014 v Ljubljani, v okviru multikonference "Informacijska družba" IS'2014. Konferenca je namenjena članom Slovenskega društva za jezikovne tehnologije (SDJT) in drugim, ki jih to področje zanima, kot forum, kjer lahko predstavijo svoje delo v preteklih dveh letih, kolikor je minilo od zadnje konference o jezikovnih tehnologijah, organizirane v okviru IS.

Zbornik vsebuje 31 prispevkov, ki obravnavajo široko paleto raziskav. Trije prispevki, od tega dva vabljena, obravnavajo raziskovalno infrastrukturo CLARIN, ki naj bi služila raziskavam s področij humanistike in družbenih ved s tem, da nudi dostop do jezikovnih virov in storitev. Več prispevkov predstavlja rezultate ali načrte evropskih in slovenskih raziskovalnih projektov. Posebno omembo si zaslužijo številni prispevki hrvaških kolegov, v katerih poročajo o izgradnji novih jezikovnih virov in o metodah strojnega učenja, ki so jih uporabili za raznovrstna jezikoslovna označevanja hrvaškega jezika. Poleg tega v zborniku najdemo tudi raziskave s področja govornih tehnologij, opise korpusnih raziskav, pregledne članke in predstavitve aplikacij.

Organizatorji bi se radi zahvalili vsem, ki so prispevali k uspehu konference: vabljenim predavateljem, avtorjem prispevkov, programskemu odboru za izjemno kvalitetno recenzentsko delo ter organizatorjem IS'2014.

Oktober 2014
Ljubljana

Tomaž Erjavec
Jerneja Žganec Gros

Preface to the Proceedings of the Ninth Language Technologies Conference

These proceedings contain the papers presented at IS-JT 2014, The Ninth Language Technologies Conference held on October 9th, 10th 2014 in Ljubljana, in the scope of the Information Society multiconference, IS'2014. The conference was aimed at the members of the Slovenian Language Technology Society, others interested in the field, as a forum where they could present their work in the last two years, which have passed since the previous IS Language Technologies Conference.

The proceedings contain 31 contributions, which present a wide variety of research topics. Three papers, of which two were invited contributions, present the CLARIN research infrastructure, which aims to facilitate research in the humanities, social sciences by enabling access to language resources, services. Several papers presents results or plans for European on national research projects. A special mention should be given to the numerous papers by our Croatian colleagues, where they report on the compilation of new language resources, machine learning methods applied to a wide spectrum of linguistic annotation tasks. The proceedings also contain descriptions of research on speech technologies, corpus linguistic research, overview papers, and presentations of applications.

The organisers would like to thank the many people who contributed to the success of the conference: the invited speakers, the authors of contributions, the programme committee for their exemplary work in reviewing the papers, and to the organising committee of IS 2014.

October 2014
Ljubljana

Tomaž Erjavec
Jerneja Žganec Gros

RECENZENTI / Program Committee

Predsednika / Chairs:

Tomaž Erjavec	Odsek za tehnologije znanja, IJS
Jerneja Žganec Gros	Alpineon, d.o.o.

Simon Dobrišek	Fakulteta za elektrotehniko, Univerza v Ljubljani
Darja Fišer	Filozofska fakulteta, Univerza v Ljubljani
Ivo Ipsić	Tehnična fakulteta, Univerza v Reki
Primož Jakopin	Inštitut za slovenski jezik, ZRC SAZU
Zdravko Kačič	Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Simon Krek	Lab. za umetno inteligenco, IJS
Cvetana Krstev	Filozofska fakulteta, Univerza v Beogradu
Nikola Ljubešić	Odsek za informacijske in komunikacijske znanosti, Univerza v Zagrebu
Nataša Logar	Fakulteta za družbene vede, Univerza v Ljubljani
Birte Lönneker-Rodman	Across Systems GmbH
France Mihelič	Fakulteta za elektrotehniko, Univerza v Ljubljani
Dunja Mladenić	Laboratorij za umetno inteligenco, IJS
Marko Stabej	Filozofska fakulteta, Univerza v Ljubljani
Tomaž Šef	Odsek za inteligentne sisteme, IJS
Jan Šnajder	Fakulteta za elektrotehniko in računalništvo, Univerza v Zagrebu
Darinka Verdonik	Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Špela Vintar	Filozofska fakulteta, Univerza v Ljubljani
Janez Žibert	Fakulteta za matematiko, naravoslovje in informacijske tehnologije, Univerza na Primorskem

User-driven Language Technology Infrastructure – the Case of CLARIN-PL

Maciej Piasecki

G4.19 Research Group
Wrocław University of Technology
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław
maciej.piasecki@pwr.edu.pl
www.clarin-pl.eu

Abstract

The paper discusses a user-driven development of CLARIN-PL, the Polish branch of the European language technology infrastructure for Humanities and Social Sciences. CLARIN-PL can be used as an exemplar of a bi-directional (i.e. top-down and bottom-up) approach to developing language resources and tools. The paper presents an overview of the state of the basic processing chain for Polish, the set of basic Polish language resources and tools and typical processing schemes emerging from the development of key applications. We also discuss the problem of the quality of services offered by language tools that goes much beyond the typical measures used during testing. In conclusion, we try to envisage further user needs and further language technology infrastructure development for which the 3-4 year construction phase is a good starting point for a fully-fledged infrastructure.

Uporabniško usmerjena jezikovnotehnološka infrastruktura – primer CLARIN-PL

Prispevek predstavi uporabniško usmerjen razvoj CLARIN-PL, poljske veje Evropske jezikovnotehnološke infrastrukture za humanistiko in družboslovje. CLARIN-PL lahko uporabimo kot primer za dvosmeren (tj. od zgoraj navzdol in od spodaj navzgor) pristop k razvoju jezikovnih virov in orodij. Prispevek poda pregled stanja osnovnega zaporedja obdelav jezikovnih podatkov, množico osnovnih jezikovnih virov za poljski jezik in orodij ter tipične sheme za obdelavo, ki izvirajo iz razvoja ključnih aplikacij. Prispevek obravnava tudi problem kakovosti storitev jezikovnih orodij, ki presega tipične ukrepe, ki se uporabljajo med testiranjem. V zaključku je podan oris nadaljnjih potreb uporabnikov in razvoja jezikovnotehnološke infrastrukture, za katerega je 3-4 letno obdobje izgradnje dobra osnova za popolnoma izdelano infrastrukturo.

1. Introduction

Language technology infrastructure (LTI) is a complex system that enables combining language tools with language resources into processing chains (or pipelines) with the help of a software framework. The processing chains are next applied to language data sources in order to obtain results interesting from the perspective of research needs of different groups of users.

Addressing user needs is the basic challenge in software engineering. Users make all systems imperfect, but the truth is that software systems do not exist without users. They simply do not have a purpose. Moreover, LTI is interesting only when its proper users are significantly different from its constructors, as it would be good to finally see language technology (LT) going beyond the level of toy systems. Basically, language engineers should not construct LTI mainly for themselves.

Users should be present at all stages of system development. In a user-driven system development process, the Context of Use[1] determines the perspective from which the users perceive LTI. Usability (defined in terms of *efficiency*, *effectiveness* and *satisfaction* (ISO, 1997 1999)) is the basis for the assessment of any interactive system including LTI.

In this paper we will discuss consequences of the user-driven development for LTI construction. We will focus on the exemplar of CLARIN – a European LTI which is meant to support researchers in Humanities and Social Sciences (H&SS). CLARIN intended users are significantly different from its constructors and usually do not possess any knowledge of computational linguistics or programming.

[1]Context of Use encompasses users and their characteristics relevant to the general goals of the future system, users' tasks and their effects and different kinds of environment (technical, organisational, social and cultural).

2. Language Technology Infrastructure

LT has been developed for more than 10 years now. LT originated from the change of small limited systems characteristic for early NLP into robust text processing technology based on sets of exchangeable and reusable components: dynamic – language tools and static – language resources.

The idea of LTI comes from the observation that we can identify several barriers that prevent wide spread use of LT outside the world of computational linguists and computer scientists, cf (Wittenburg et al., 2010), namely :

- *physical* – language tools and resources are not accessible in the network,
- *informational* – descriptions are not available or there is no means for searching,
- *technological* – lack of commonly accepted standards for LT, lack of a common platform, varieties of technological solutions, insufficient users' computers,
- related to *knowledge* – the use of LT requires programming skills or knowledge from the area of natural language engineering,
- *legal* – licences for language resources and tools (LRTs) limit their applications.

LTI is a complex system providing a technological platform for the integration of different LT components into

one interoperable system. Moreover, other aspects like legal and informational ones are also taken into account.

CLARIN is a LTI focused on the use in the area of H&SS. The main goal of CLARIN is to decrease the barriers, as far as possible in the context in which LT is used by researchers from H&SS.

3. Development Schemes

CLARIN[2] is being built by an ERIC consortium of several countries that are obliged to contribute parts of the LTI. Different countries follow different schemes, however some common features can be identified. There are two possible basic schemes. The first is a *bottom-up process*, which can be also termed a *collected offer*. It is based on linking the already existing LTRs, and it is focused on establishing accessibility and technical interoperability of LTRs, as well as on establishing a common system of IPR licences that lowers the legal barrier. A distributed authorisation system is introduced and federated search mechanisms for searching the content of the resources and metadata in pre-defined formats (Wittenburg et al., 2010) are proposed. As a result, the tools and resources will become accessible via Web to the users and can be combined into processing chains. The only question left open is if the users know what to look for and what to use. LRTs mostly require from users the specific background knowledge, e.g. complex Slavic tagsets. LRTs often seem not to be directly related to the research performed by the users from H&SS.

Web applications both for individual services and for adaptable workflows for natural language processing for final users are mostly on borders of the main focus of CLARIN. *Usability aspects* (ISO, 1997 1999) and especially *usability evaluation* of the applications are very often neglected. At the same time, data presentation in resources and the results of processing in tools are implemented according to the user needs that are unknown! Processing chains are adapted to the unknown user tasks, whose goals mostly go beyond the domain of natural language engineering. However, at the same time, LTI is a new enabling technology that can create new needs, if well presented and explained. Sample applications that illustrate the possibilities on real examples can be very important tool in this task and can potentially inspire the future users.

The second possible, but probably never thoroughly implemented approach is based on the *user-centred design* paradigm (Hackos and Redish, 1998). It can be called a *top-down process*, as the starting point are complete research applications (or research tools) for the final users – H&SS scientists. Requirements for the applications should be discovered by applying methods of Context of Use Analysis. Next, research applications and the underlying network system of services and LT components should be designed and developed according to the requirements.

Despite the expected large number of LRTs that can be immediately re-used in the constructed infrastructure, this approach seem to be unrealistic. Research tools to be designed are innovative and are associated with the development of new research methods. Their discovery could be much easier through working prototypes and experiments for selected limited subdomains of H&SS. A long way from the project to the results and the perspective of costly long term investment could be unacceptable.

In comparison to the pure user-centred approach, a mixed option of a bi-directional process seems to be more practical. According to this approach, the existing LTRs, possibly many, are combined into a distributed network infrastructure, too. However, user-driven requirements are also taken into account. Designing the top level research applications for users is a starting point for many activities in the LTI development. The infrastructure construction process follows a metaphor of the Agile-like (Larman, 2004) light weight software designing method. Key users are identified and prototype research applications are created in co-operation with them and according to the requirements acquired from them. The application development stimulates the construction of technical fundaments, and inspires the identification of further user needs on the basis of analogy to the working prototypes.

4. Bi-directional approach of CLARIN-PL

In spite of significant improvement that had been made in the area of LT for Polish since 2005, quite many basic LTRs for Polish were still lacking at the start of CLARIN-PL (Jan. 2012). This situation resulted in deepening the technological barrier, as LRTS necessary for many applications simply did not exist. One of the most typical examples is the lack of a robust dependency parser for Polish – many application for English take the existence of such a parser for granted. Thus, the target CLARIN-PL structure is based on three parts:

1. CLARIN-PL Language Technology Centre[3] – the Polish node of the CLARIN distributed infrastructure,
2. a complete set of basic LRTs for Polish,
3. research applications for H&SS – first created for key users and selected H&SS sub-domains.

The LT centre is meant to provide fundamental CLARIN facilities (Roorda et al., 2009) like distributed authorisation and archiving system for LT supporting the CMDI meta-data format (Broeder et al., 2009) and persistent identifiers. A special focus is given to collecting LRTs for Polish and making them accessible via web services and linking them into processing chains. Moreover, the web services are accompanied by web-based applications with user interface in Polish[4]. A CLARIN centre with such functionality is classified as a CLARIN B-type centre (Roorda et al., 2009). As the number of resources (both text and speech) is limited among the CLARIN-PL partners, we plan to build interfaces linking the LT centre with the existing archives and repositories, e.g. digital libraries, and with other research infrastructures, e.g. DARIAH. However, the ongoing process of distributing the workload inside CLARIN ERIC causes that the Polish centre also plans

[2] www.clarin.eu

[3] http://www.clarin-pl.eu

[4] This requirement is very important, as many users from H&SS do not accept user interface in English.

to take responsibility for selected services that are fundamental for the whole infrastructure, i.e. elements of the responsibility of the CLARIN A-type centre.

A starting point for the identification of the missing LRTs for Polish was the comparison of the list of LRTs for Polish available on open licences with the BLARK[5] set of LRTs (Krauwer, 1998; Krauwer, 2003), as well as with the basic processing chains of Information Extraction. We envisaged the latter as the most likely scheme for applications. BLARK was selected as a *quasi* standard or standard *de facto* as the set of LRTs proposed in BLARK has already been a target point in development of LT for several languages. We assumed that implementation for Polish of all LRTs assumed in the BLARK set would increase interoperability between CLARIN-PL and the rest of CLARIN infrastructure, as the BLARK set is often used as a reference point. The bare existence of LRTs does not mean that they fit to CLARIN needs. Applications in H&SS impose high demands on their coverage and quality. Several of the existing LRTs must to be significantly expanded in CLARIN-PL in order to make them robust enough, see Sec. 5.

The multilinguality of CLARIN infrastructure makes the construction of bilingual resources crucial for interoperability. Balancing between the coverage and a range of resource types, we decided to concentrate mainly on bilingual Polish-English resources. An overview of LRTs planned to be developed in CLARIN-PL is presented in Sec. 5.

In a similar way to other CLARIN national consortia, general and flexible work-flows have to be constructed in order to facilitate the full use of different language tools.

Digital H&SS (also e-Humanities and e-Social Science) are developing very quickly, but they are still relatively new domains with not many fixed procedures. Research tasks in these domains are approached in a very dynamic way with a rich variety of specific solutions based on decisions dependent on research data and interim results. There are several methods for gathering informations about the Context of Use proposed in Human Computer Interaction (Hackos and Redish, 1998), but *direct observation* is considered to be the best one as it allows for observing users performing their tasks in their natural environment. In our case, the users are researchers and they perform scientific research. Users participating in the observation sessions should be representative. However, if the designer does not possess deep knowledge about the given domain, which still seems to be the case of Digital H&SS, the first group of the users, called *key users*, can be composed from those who are somehow characteristic to the domain. As key users, we selected scientists from H&SS who have already started using digital language-based methods in their research or are interested in applying LT in their research.

[5]BLARK is the acronym for *The Basic Language Resource Kit* and it is "the minimal set of language resources that is necessary to do any precompetitive research and education at all" (Krauwer, 2003). A BLARK comprises different kind of resources and tools treated as a minimal required set for every language. This quasi-standard has been implemented for several languages and become a main reference point for evaluation of the state-of-the-art of the LT for a particular language.

Direct observation is mainly based on collaboration with key users in their research projects. After collecting information about the Context of Use, possible LT-based techniques that can support the research are selected or even a new research process is defined. The method consists of several steps:

1. Establishing contacts with users
2. Identification of key users
3. Context of Use Analysis: users, their tasks and environments
4. Identification of the key applications corresponding to these users' tasks that can be supported by the available LT.

The first contacts with the prospective key users were established on the basis of the previous personal acquaintance with particular H&SS researchers. These direct links resulted in our participation in a couple of H&SS conferences and further contacts. Direct communication with possibly many conference participants appeared to be fruitful. Most key users are researchers who have already started using or are interested in using computer system in their research.

From the very beginning of the project we have been using our CLARIN-PL web page to inform potential users about the project. We published a list of of generally described potential CLARIN LTI applications with a special focus given to Polish LT. Our intention was to make H&SS researchers aware about existing possibilities and also to associate the CLARIN-PL web page with potential topics searched on the Web. We try to keep the list constantly growing, e.g. on the basis of the experiences collected during the application development. Moreover, on the CLARIN-PL web page portal we have also started collecting information about Polish conferences and projects from the domain of Digital H&SS and related domains – valuable information is the best advertisement on the Web.

The established contacts with the prospective users allowed us to select the first group of key applications discussed in Sec. 6. We tried to cover a maximal variety of research areas, but also to co-operate first with the most active users. During this first round, the number of applications is a less important factor, and we had aimed at only a few, e.g. due to the financial limitations. The available LT for Polish was also a limiting factor in this selection. We assumed that the first constructed applications would significantly broaden our understanding of the domain and help to identify further application domains or even generalise the constructed applications to general frameworks.

As a result, CLARIN-PL can be treated as an exemplar of a bi-directional approach combining together bottom-up and top-dow development. We are trying to harmonise these two approaches, i.e. to interactively shape the LRT development plan according to requirements collected from the work on key applications, e.g. CLARIN-PL tasks in the area of Information Extraction have been re-organised and re-ranked due to the collected experience. Summing up, the bi-directional approach is a fruitful scenario: key users,

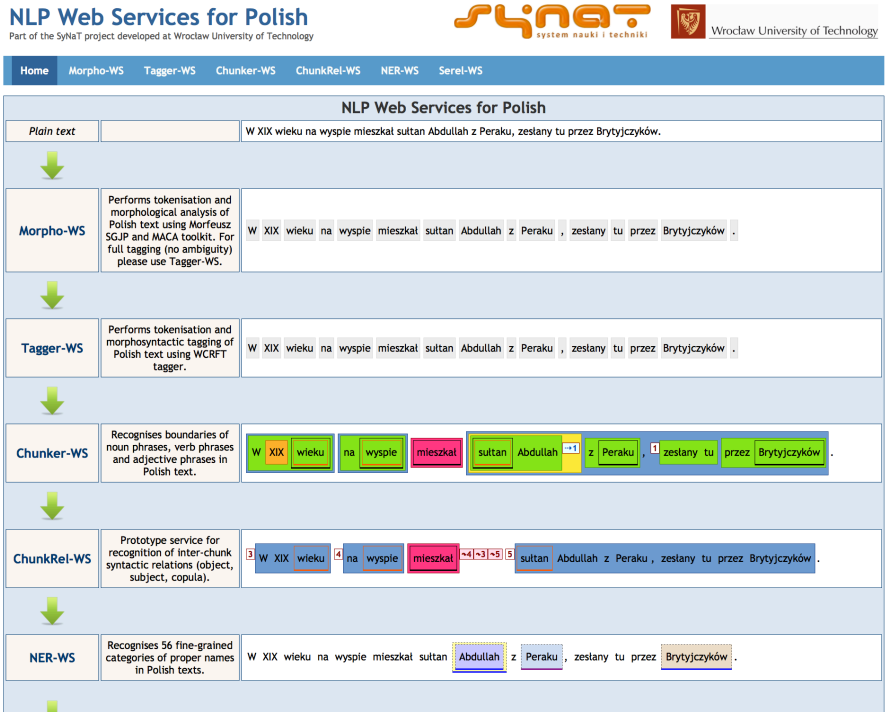


Figure 1: A prototype of the CLARIN-PL basic processing chain for Polish.

research tasks identified, key applications can provide generalisations resulting in adaptable research workflows.

5. Resources and Tools for Polish

The necessity of the significant improvement of basic LTRs for Polish was identified as a pre-requisite for lowering the technological barrier. On the basis of the analysis of the state-of-the-art of LT for Polish done at the start of CLARIN-PL, we planned several of tasks within this goal.

5.1. Resources and Supporting Technology

Concerning language resources, the starting point of CLARIN-PL was relatively good as several basic resources had been constructed and become matured, e.g. a huge National Corpus of Polish (Przepiórkowski et al., 2012), very large Polish wordnet – plWordNet (Maziarz et al., 2013) and an open KPWr corpus of Polish[6] with rich annotation (Broda et al., 2012). Thus, our main goals are completing the construction of selected resources and building bi-lingual resources and specialised corpora facilitating the envisaged needs of H&SS.

We plan to expand plWordNet to a comprehensive description of the Polish lexico-semantic system (with around 260 000 lexical units) and fully map it to Princeton WordNet 3.1 (Fellbaum, 1998)[7]. A large lexicon of the Multi-word Expressions described with the minimal constraints on their lexico-syntactic structures (Kurc et al., 2012) will be expanded up to the size of 60 000 Polish Multi-word Expressions manually described. All of them will be semantically described in plWordNet 3.0. The lexicon of semantically classified proper names (*NELexicon*[8]) will be expanded to 2.5 million distinct PNs. For both lexicons we will construct robust tools for their further automated expansion on the basis of corpora. This is meant to be an implementation of the idea of a dynamic lexicon, i.e. a combination of the core described manually and a large part extracted automatically from selected corpora on demand of the user. The automated tools will allow the CLARIN users to create their own domain extensions of both lexicons. The users will be also equipped in editors for the manual verification of the automatically extracted data. A large semantic valency lexicon for Polish predicative lexical units (verbs, nouns) will be also constructed (Hajnicz, 2014). Semantic restrictions on valency frame arguments will be described be described by means of the selected plWordNet synsets that are more general and define hypernymy subhierarchies used as represent ions of semantic domains.

Concerning corpora, CLARIN-PL is going to build: a transcribed training-testing Polish speech corpus, a corpus of Polish conversational texts transcribed from speech recordings and annotated parallel corpora mapping Polish text to several languages (Bulgarian, Russian and Lithuanian). Historical Polish corpus of text news from 1945-1954 that will be also developed is a resource directly fo-

[6] *Korpus Politechniki Wrocławskiej* (Wrocław Univeristy of Technology Corpus, `http://nlp.pwr.wroc.pl/kpwr`) in an open corpus of Polish which is balanced according to different genres and built from texts on Creative Commons. Currently, KPWr includes 449 000 tokens of text documents of 5 styles. KPWr has been annotated on several different levels of the linguistic structure, e.g. shallow syntactic structures (161 716 chunk annotations and), proper names, anaphora, semantic relations etc.

[7] As WordNet 3.1 appeared to be too small for providing mapping targets for all Polish senses we have initiated a significant expansion of WordNet 3.1 as a part of the CLARIN-PL plan.

[8] `http://nlp.pwr.wroc.pl/nelexicon`

cused on applications in H&SS.

In order to fully utilise the rich set of corpora, several systems for searching text and speech corpora will be expanded or built. A system for semantic indexing of large text corpora on the basis of publicly available encyclopaedias will be built.

5.2. Tools

The situation of the basic processing chain for Polish at the beginning of CLARIN-PL is presented below. Robust tools are presented in bold. Tools existing in prototypes with limited accuracy and coverage are written in normal font, and non-existing tools are shown in italic font.

1. **Segmentation into tokens and sentences.**
2. **Morphological analysis.**
3. Morphological guessing of unknown words (both without context and context sensitive).
4. **Morpho-syntactic tagging**.
5. Word Sense Disambiguation.
6. Chunker and shallow syntactic parser.
7. Named Entity Recognition and disambiguation.
8. Co-reference and anaphora resolution.
9. Temporal expression recognition.
10. Semantic relation recognition.
11. Event recognition.
12. *Shallow semantic parser.*
13. Deep syntactic parser *with disambiguated output*: dependency and *constituent*.
14. *Deep semantic parser.*

As most Polish language tools have been constructed with the focus on standard language and error-free text, an important element of the plan is the construction of a generic set of morpho-syntactic tools for Polish that can be adapted to a domain specified by the user.

We also plan to work on tools for the extraction of the semantic-pragmatic information from documents and collections of documents (e.g. keywords, semantic relations between text fragments and text summaries) and an open stylometric and textometric system.

All language tools presented above are used by CLARIN-PL or will be expanded or developed by CLARIN-PL. We plan to provide web services for all of them and also to include them into the processing chain. By now, we have implemented web services for: segmentation, morphological analysis, tagging, chunker and Named Entity Recognition[9]. Prototype web services for Word Sense Disambiguation and Semantic relation recognition are ready, but their accuracy is not yet satisfactory. There is also a web service providing access to plWordNet 2.2. Web services are accessible via both REST and SOAP and their programming interface is specified in WSDL language. We plan to describe them in CMDI meta-data format and integrate with the repository system of CLARIN-PL Language Technology Centre.

A prototype of the user interface for an implementation of the basic processing chain is presented in Fig. 1. The whole chain and its components are available as web services described with meta-data in CLARIN CMDI format. We plan to link them to WebLicht platform (Hinrichs et al., 2010) and also to build our own platform for defining processing chains focused on Polish users.

6. Research Applications

Only programs or systems constructed in response to real CLARIN users' (i.e. H&SS researchers) requests can be treated as CLARIN applications. Interactive systems that are not used do not exist.

A search system for the corpus of conversational data *Spokes*[10] was constructed in the close co-operation with linguists inside the CLARIN-PL consortium. So, it is not a genuine application, but it provides rich facilities for not only searching the corpus, but also for statistical analysis of the retrieved data. Corpus search tools are basic application that mostly provide only searching through language data, but anyway they are crucial applications. However, the issues of rich annotation, big data volumes and statistical analysis of the query results, the construction of the corpus search tools is much more challenging.

Requests from users sometimes reveal gaps in the available technology that were not expected before the project start. Several tools for web-based corpus building appeared to be too sensitive to text encoding errors found in the web (e.g. a different code page declared in meta-data than really used). As a result a system for collecting Polish text corpora from the Web had to be constructed. The system is combined with morphological analysis in order to detect texts including larger number of errors (or non-words). The system was also requested to provide support for semi-automated extraction from blogs only those elements that fit to the pre-defined user requirements.

There were several textometric and stylometric tools available, but none of them was well suited for rich inflection of Polish, e.g. the available tools did not provide support for lemmatisation and tagging of Polish. We plan to build a system for Polish enabling the use of features defined on any level of the linguistic structure: from the level of word forms up to the level of the semantic-pragmatic structures. The system will combine several existing components: language tools for pre-processing, *Fextor* (Broda et al., 2013) – a system for defining features in a flexible way, *Stylo*[11] – a stylometric package for English, *SuperMatrix* (Broda and Piasecki, 2013) – a system for building and processing very large co-incidence matrices with linking to clustering and Machine Learning packages.

[9] The services are available at `www.clarin-pl.eu/en/services/`

[10] `http://clarin.pelcra.pl/Spokes/`

[11] `http://crantastic.org/packages/stylo/versions/34587`

Stylometric techniques appear to be applicable in many tasks of H&SS that are based on the comparison of texts, e.g. in sociology (features that are characteristic for different subgroups), political studies (similarity and differences between political parties), literary studies (analysis of blogs as creative work types), etc. The extended system will allow for the analysis of the semantic associations of words on the basis of Distributional Semantics, semantic relation extraction, collocations, semantic comparison of texts and texts collections, etc. Thus, it starts to be similar to a system for the semantic text classification discussed in the next subsection.

6.1. Semantic Text Classification for Sociology

One of our first research application is an exemplar of the scheme which can be generalised to many projects in H&SS. One of the goals of the research project realised in *Collegium Civitas* (a non-state university) in Warsaw was to check the content of web pages of the Polish institutions (public and private) related to culture, in its broadest sense. Around 3200 institution were pre-selected and almost 200 000 documents were acquired from their web sites. The content of the web pages was divided into paragraphs of different sizes (around 1 200 000). The goal was to classify the paragraphs into 20 semantic classes defined by the sociologists. The classes describe different aspects of the use of the web page as a communication medium and they were organised into three groups: competences, functions of the culture, thematic areas plus 6 individual classes (e.g. auto-presentation or local function).

The initial vision was a simple system for supervised classification of text documents. After the Context of Use Analysis, the plan was expanded to a complex system encompassing user-controlled corpus building, text preprocessing (text segmentation and morpho-syntactic tagging and parsing), automated sample selection, manual annotation, training classifiers and automated annotation and result analysis. Moreover, we discovered that there is no open corpus annotation editor focused on applications in Social Sciences. The constructed prototype system can be also adopted to many similar tasks in Digital H&SS.

6.2. Literary Map

Literary Map is a CLARIN-PL application that has originated from a concrete user request formulated during an open part of a CLARIN-PL working meeting. The first prototype is presented in Fig. 2. The user is Digital Humanities Centre of The Institute of Literary Research of PAS. The main idea is to identify all geographical names in the literary text (or a corpus) and map them onto the geographical map. The task goes beyond Named Entity Recognition (NER), as NER must be combined with geo-location. We use geo-location service provided by Google, but still location PNs recognised in text must be grouped into expression recognised by Google in a way enabling good accuracy of locating them. We proposed to expand the initial idea with recognition of semantic relations linking non-spational PNs in the text with the location PNs and visualising those links on the map, too. Recognition of the temporal expression could further enrich the application.

Two scenarios of use are considered: fully automated and bootstrapping. According to the first, users process whole corpora of literary texts and next can analyse collected statistical data or browse mapping of the individual texts. However, due to the limited accuracy of the whole system, the second scenarios in which the system is used a supporting tool during corpus annotation with mappings is more likely in research – annotations proposed by the system are next corrected by the researchers.

7. Conclusions

The Quality of Service notion is very rarely used in relation to LRTs and LTI. However, this is the crucial question: for what research tasks and what scenarios are our LRTs good enough? If we aim at fully automated procedures, the expected quality is very high, e.g. 5% can bias a lot statistical analysis of data extracted from a corpus. Application of LT to the research in H&SS seem to be much more challenging than in commercial systems! We need to develop a model of LT-based applications in which we can describe and manage errors introduced by different LRTs and their accumulated influence on the final result of the whole application.

Semi-automated model in which LT-based applications are used for preparing initial text annotation, next corrected by researchers, or supporting researchers in browsing corpora and finding examples is the most likely way. Here, Visualisation of the results on different stages comes into play as a very important element of LTI.

Any model of LTI we aim for users should be the starting point for LTI development and also the goal for this work.

Acknowledgement: The work was financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education and was supported by the EU's 7FP under grant agreement no 316097 [ENGINE].

8. References

Bartosz Broda and Maciej Piasecki. 2013. Parallel, massive processing in SuperMatrix – a general tool for distributional semantic analysis of corpora. *International Journal of Data Mining, Modelling and Management*, 5(1):1–19.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. 2012. KPWr: Towards a Free Corpus of Polish. In N. Calzolari et al., editor, *Proc. of the Inter. Conference on Language Resources and Evaluation, LREC'12*, Turkey. ELRA.

Bartosz Broda, Paweł Kędzia, Michał Marcińczuk, Adam Radziszewski, Radosław Ramocki, and Adam Wardyński. 2013. Fextor: A Feature Extraction Framework for Natural Language Processing: A Case Study in Word Sense Disambiguation, Relation Recognition and Anaphora Resolution. In Adam Przepiórkowski, Maciej Piasecki, Krzysztof Jassem, and Piotr Fuglewicz, editors, *Computational Linguistics*, volume 458 of *Studies in Computational Intelligence*, pages 41–62. Springer Berlin Heidelberg.

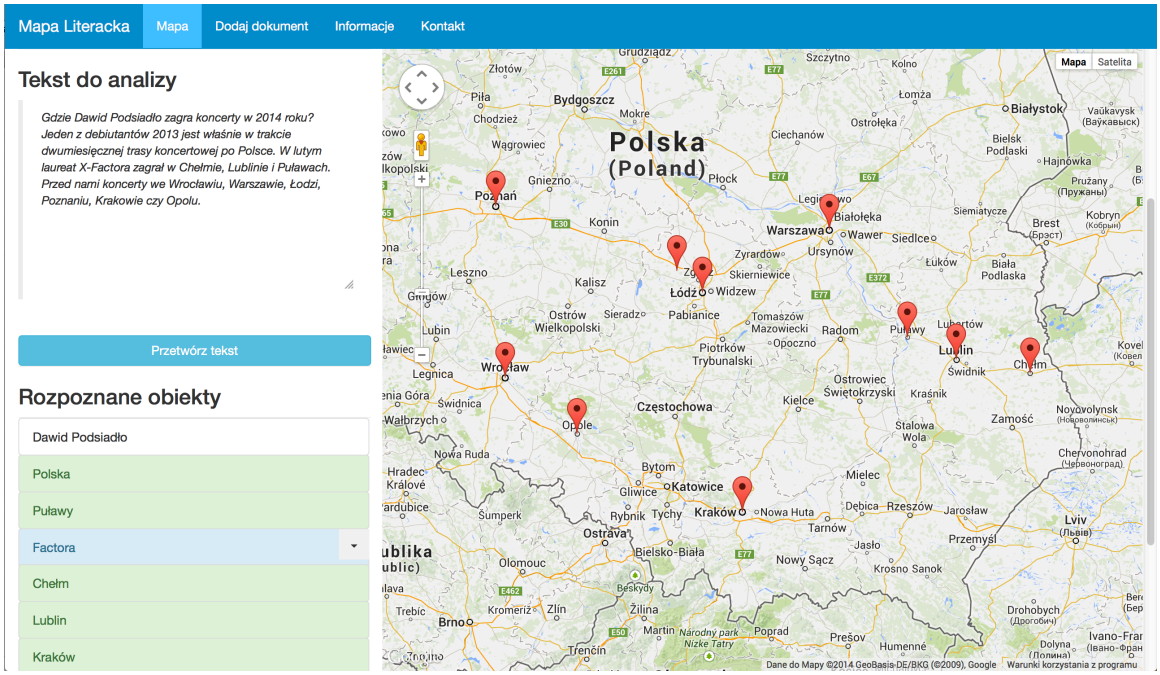


Figure 2: A prototype of the Literary Map application.

Daan Broeder, Bertrand Gaiffe, Maria Gavrilidou, Erhard Hinrichs, Lothar Lemnitzer, Dieter van Uytvanck, Andreas Witt, and Peter Wittenburg. 2009. Registry requirements metadata infrastructure for language resources and technology. Technical Report CLARIN-2008-5, Consortium CLARIN. PID: `http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-33`.

Christiane Fellbaum, editor. 1998. *WordNet — An Electronic Lexical Database*. The MIT Press.

J. Hackos and J. Redish. 1998. *User and Task Analysis for Interface Design*. Wiley Comp. Pub.

Elżbieta Hajnicz. 2014. Lexico-semantic annotation of *składnica* treebank by means of PLWN lexical units. In Orav et al. (Orav et al., 2014), pages 23–31.

Erhard Hinrichs, Marie Hinrichs, and Thomas Zastrow. 2010. WebLicht: Web-based lrt services for german. In *Proceedings of the ACL 2010 System Demonstrations*, ACLDemos '10, pages 25–29, Stroudsburg, PA, USA. Association for Computational Linguistics.

ISO. 1997–1999. *ISO 9241 — Ergonomic Requirements for Office Work with Visual Display Terminals*. ISO.

Steven Krauwer. 1998. BLARK: The Basic Language Resource Kit. ELSNET and ELRA: Common past, common future. Web page. Accessed 16th Sep. 2014.

Steven Krauwer. 2003. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. In *Proceedings of SPECOM 2003*.

Roman Kurc, Maciej Piasecki, and Bartosz Broda. 2012. Constraint based description of polish multiword expressions. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2408–2413, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Craig Larman. 2004. *Agile and Iterative Development: A Manager's Guide*. Addison-Wesley.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2013. Beyond the transfer-and-merge wordnet construction: plWordNet and a comparison with WordNet. In *Proc. of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 443–452, Hissar, Bulgaria. INCOMA Ltd.

Heili Orav, Christiane Fellbaum, and Piek Vossen, editors. 2014. *Proceedings of the 7th International WordNet Conference (GWC 2014)*, Tartu, Estonia. University of Tartu.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warszawa.

Dirk Roorda, Dieter van Uytvanck, Peter Wittenburg, and Martin Wynne. 2009. Centres network formation. Technical report CLARIN-2008-3, Consortium CLARIN. PID: `http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-27`.

Peter Wittenburg, Núria Bel, Lars Borin, Gerhard Budin, Nicoletta Calzolari, Eva Hajicová, Kimmo Koskenniemi, Lothar Lemnitzer, Bente Maegaard, Maciej Piasecki, Jean-Marie Pierrel, Stelios Piperidis, Inguna Skadina, Dan Tufis, Remco van Veenendaal, Tamás Váradi, and Martin Wynne. 2010. Resource and service centres as the backbone for a sustainable service infrastructure. In N. Calzolari et al., editor, *Proc. of the International Conference on Language Resources and Evaluation, LREC 2010, Malta*, pages 60–63. ELRA.

CLARIN-DARIAH.AT - Weaving the network

Matej Ďurčo, Karlheinz Mörth
Austrian Centre for Digital Humanities, Austrian Academy of Sciences
Vienna, Austria
E-mail: matej.durco@oeaw.ac.at, karlheinz.moerth@oeaw.ac.at

Abstract

The paper gives an overview of recent developments in Austria regarding CLARIN and DARIAH, the two Digital Humanities research infrastructure consortia. The presentation touches on the manifold related international engagements as well as a new wave of national activities and projects. Special attention is directed towards semantic technologies which are becoming the focal point of a diverse range of research areas in the digital humanities, raising the question how HLT can support other disciplines to cope with the "semantic turn".

CLARIN-DARIAH.AT – spletanje omrežja

Prispevek poda pregled nedavnega razvoja konzorcijev raziskovalnih infrastruktur za področje digitalne humanistike CLARIN in DARIAH v Avstriji. Predstavljene so številne mednarodne povezave, kot tudi novi val nacionalnih aktivnosti in projektov. Posebna pozornost je namenjena semantičnim tehnologijam, ki postajajo žarišče raznovrstnih področij raziskav v digitalni humanistiki, s čimer se sproža vprašanje, kako lahko jezikovne tehnologije podpirajo ostale discipline in se spopadejo s »semantičnim obratom«.

Keywords: research infrastructures, digital humanities, semantic technologies

1. Looking back

Austria has been involved in CLARIN[1] and DARIAH[2] already since 2009. It actively contributed to the build-up of technical infrastructures and engaging in the set-up of the organisational structures. At that time, the main contributors were the Centre for Translation Studies at the University of Vienna, the Institute for Corpus Linguistics and Text Technology of the Austrian Academy of Sciences (ICLTT) and the Centre for Information Modelling, University Graz (ZIM). Within CLARIN, the contributions were chiefly related (a) to the Component Metadata Infrastructure (Broeder et al., 2010) – the prototypical development of individual exploitation-side modules, especially the Semantic Mapping Component (Ďurčo, 2013) – as well as (b) to the FCS – Federated Content Search, an initiative aiming at developing a distributed system allowing to search not only in metadata, but also in the content of the resources exposed by individual data providers (Stehouwer et al. 2012).

Even before the beginning of the pan-European research infrastructures the ICLTT (and its predecessor the Austrian Academy Corpus – AAC) had a long tradition in Digital Humanities dating back to the late 1990s. The most prominent example may be the AAC-FACKEL, the digital scholarly edition of the magazine “Die Fackel”[3] authored by Karl Kraus in the years 1899 until 1936. The institute looks also back on a tradition of experimental dictionary-making. More recently, both monolingual and bilingual lexicography has again gained in importance. The focus in these efforts has been on developing tools, working on lexicographic data and eLexicography standards.

Equally, ZIM has a long tradition in conducting DH projects, mostly through the well-proven strategy of accompanying humanities projects, offering them the expertise regarding data modelling, preservation and online publication. The technical heart of these activities is a fedora-based repository system called GAMS[4], which has been developed there since 2003.

These are only a few examples of digital humanities related research in Austria. A survey conducted in 2009 listed more than 30 projects producing digital language resources by 15 different research groups. However most activities have been performed as solitary projects not embedded in any larger framework. Much work needs to be done to achieve a higher degree of integration.

2. New phase – Austrian Centre for Digital Humanities

In 2013, the broad range of CLARIN and DARIAH activities carried out in the last years were brought together in a new initiative, the Austrian Centre for Digital Humanities / Digital Humanities Austria (DHA), a project being funded by the Ministry of Science, Research and Economy for the duration 3 years. Digital Humanities Austria has been designed as a platform and a network of excellence for the propagation and dissemination of the digital paradigm and the use of DH methods and technologies. DHA represents the Austrian implementation of the EU’s ESFRI[5] roadmap.

Working with digital resources and tools remains a methodological and logistical challenge for many researchers in the humanities. DHA fosters the use of digital data, tools and know-how by easing access, enabling

[1] Common Language Resources and Technology Infrastructure http://clarin.eu/
[2] Digital Research Infrastructure for the Arts and Humanities http://dariah.eu/
[3] http://corpus1.aac.ac.at/fackel
[4] http://gams.uni-graz.at
[5] European Strategy Forum on Research Infrastructures

the production of standards-based data and deepening existing skills. All these activities are conducted in close cooperation with institutes of the Austrian Academy of Sciences, the Austrian universities as well as other institutions in the country that conduct or support relevant research such as libraries, museums, archives etc.

2.1 Central concerns

Digital Humanities Austria is based on a new understanding of scholarly research which may not be reduced to simply using digitised materials. The existence of digital data is a prerequisite for digital research, however it is only one of many aspects of DH. The following principles have been agreed upon by many representatives of the digital paradigm as seminal characteristics of the new inventory of methods: systematic use of digital infrastructures, transdisciplinarity, collaborative work, participatory technologies (virtual research environments, web-based research portals), Open Access / Open Source and open life cycle of research data and research results.
The central concerns of DHA can be summarised in three key phrases explained further below:

- Save the Data
- From Data to Knowledge
- The right Toolbox

Save the Data covers all the issues related to ensuring long-term preservation and availability of existing and newly created digital research data. In many meetings with representatives of research groups at the Academy and other institutions this has been identified as a central and urgent concern. More often than not, research material produced during projects ends up undocumented on external drives and is lost for future research.
This issue has many facets, quality and format of the data including information about the data (metadata) being one, but also the availability of stable, reliable institutional or national repositories and a pressure or guidance from the funding agencies, to name the most relevant aspects.
One precondition for long-term preservation is the question of standardised formats, when modelling research data. Using standards and de-facto standards makes it far more likely that the research data can be reused by others and is compatible with external third-party systems. Consequently, the overall strategy of DHA revolves around the triad: data, tools and standards, standards representing the glue between data and tools. The preferred/default format for text-based data in DHA – which is in line with widespread usage in the DH community – is the de-facto standard TEI/XML, however it is clear that no one format can cover the diversity encountered in the broad field of DH. To tackle the issue of standardisation, both CLARIN and DARIAH have established bodies responsible for surveying existing practices and working on recommendations and guidelines.
Another aspect of data preservation are dedicated institutional repositories as crucial infrastructure components that are able to handle not just scientific publications (usually documents in PDF format), but also complex structured research data. A good example of such a repository is the GAMS run by ZIM (Graz) that offers an integrated single-sourced, multi-view system relying on community based standards.
A new addition to the repositories landscape and a major achievement of 2014 is the new CLARIN Centre Vienna[6], the first Austrian node in the network of CLARIN Centres[7] which has acquired after a comprehensive assessment procedure the DSA[8] (Data Seal of Approval) and the CLARIN Centre B status as of April 2014. The core of the CCV is the Language Resources Portal, a depositing and publishing service primarily intended for digital language resources with a humanities background (Budin et al., 2013).
The issue of long-term preservation and availability of research data has gained increased importance as funding agencies have become aware of this issue and started to demand strategies for the availability of research results and research data.
In order to raise the general awareness and to intensify the discussion about this issue, a workshop on long-term preservation of data and repositories will be held during the Austrian Days of Digital Humanities in Vienna beginning December this year. In this workshop data producers and providers of scientific repositories shall come together to discuss problems related to data management and possible solutions of these problems.
From Data to Knowledge is the second focus of DHA. Scholarly work in DH often means to enrich data, to interpret, to annotate (semantically) and to interlink data. In the build-up of a modern network of knowledge semantic approaches and the paradigm of Linked (Open) Data are expected to play central roles. (We elaborate on this further in chapter 4.)
The goal of providing *the right toolbox* for DH poses quite a challenge considering the great number of involved disciplines with their quite varied traditions and methods, their often very particular research questions that often require highly specialised digital tools. In most fields, we have not yet out-of-the-box solutions. The development and propagation of innovative tools for the digital era, so-called dedicated applications, is a major focus of our efforts to support the continuously growing number of scholars pursuing digital research. Virtual research environments that enable researchers to work collaboratively are one such type of infrastructure components.
The ICLTT has been developing two suites of tools, one being a virtual research environment for lexicographic work, the other one is a platform for online publication of digital editions, called *corpus_shell*. This framewok is developed in collaboration with Telota[9] – the technical group at the Berlin-Brandenburg Academy of Sciences and Humanities.

[6] http://clarin.oeaw.ac.at/ccv
[7] http://www.clarin.eu/centres
[8] https://assessment.datasealofapproval.org/assessment_121/seal/html/
[9] http://www.bbaw.de/telota

As mentioned before, ZIM has been developing an integrated fedora-based repository system (GAMS) that comes together with a versatile open-source client for the management of the data in the repository which also includes batch editing of data.A representative of ZIM also contributes to DARIAH as task leader for "Reference software packages" in DARIAH's Virtual Competency Centre I (eInfrastructure), inventarizing existing software usable by DH research teams.

2.2 Organizational setup

The virtual network is organized in a national consortium comprising, a number of Austrian academic institutions. Next to the core members of the preparatory phase who still ensure the continuity of activities, the Technical University Vienna, the University of Innsbruck and additional new institutes of the Austrian Academy of Sciences, the University of Vienna and the University Graz joined the consortium. The network is funded by the Federal Ministry of Science, Research and Economy and coordinated by the Austrian Academy of Sciences.

2.3 Work packages

The activities of the new initiative are organised in three main thematic areas: Research Infrastructures for Digital Humanities (*RI4DH*) which chiefly perpetuates the involvement in CLARIN and DARIAH, a digitization initiative *go!digital*, and a bundle of activities to strengthen the DH in the education *dh-curriculum*.

RI4DH comprises the technical aspects of building research infrastructures and the various engagements in the European Research Infrastructure Consortia [10] CLARIN and DARIAH, as well as the coordination of the national efforts with the respective institutions on the European level.

The overall goal is to strengthen inter- and trans-disciplinary research and development in the humanities, on the basis of European research infrastructures DARIAH and CLARIN implementing the ESFRI roadmap. This also includes the construction of a research and service platform for the collaborative work of Austria DARIAH and CLARIN partners and their operational embedding in the two ERICs.

The main part of the RI related work lies in the procurement of so-called in-kind contributions, contracted between the national consortium and the ERICs on an annual basis. For CLARIN, the contributions consist mainly in digital language resources, but also other data (such as controlled vocabularies), software packages (e.g. lexicographic tools), services (like the establishment of the Language Resources Portal) or international events (workshops, conferences). The DHd conference[11] – Digital Humanities in German-speaking area – organized by the University of Graz in February 2015 is an example of a major event as an in-kind contribution.

One source for potential new Austrian in-kinds is the Language Resources Survey conducted back in 2009 together with its update planned for this year.

go!digital was set up as a call for innovative digitization projects in Austria. 5 projects were selected by an international jury out of 36 submissions. The projects have a duration of 1,5 - 2 years, and dispose of a budget of roughly 100.000 EUR each. The call put a strong emphasis on the use of standards and the integration with research infrastructures. The high number of proposals shows also the high potential in the Austrian research landscape.

Together with a related call "Digital cultural heritage" for proejects based at the AAS (with longer duration of projects and higher volume) 10 new DH projects will start by the end of the year, which constitutes an unprecedented surge of coordinated activities in this area in Austria (though the calls were explicitly inspired by similar setups in Netherlands and Germany). Under the motto *innovation*10* all the new projects will be presented in a kick-off event on 1 December 2014 as part of the Austrian Days of Digital Humanities, organised to foster collaboration and exchange among the projects and also in the broader community.

The third area of action is education. The DHA initiative *dh-curriculum* has been motivated by the evident lack of young DH experts in the country. The initiative's particular concern are consciousness raising activities and the training of young researchers. In addition to workshops, seminars and summer schools, participating researchers work on a DH curriculum, which aims to ensure the anchoring of related know-how in the academic education. Specialized courses are supposed to enable so-called data scientists to work in their respective disciplines, to support DH projects and to curate digital data collections (corpora, editions, digital archives, etc.).

While the University of Graz already offers a complete module on DH practices, there is nothing comparable to be found in the rest of the country. However, this deplorable state of affairs is going to be changed as a number of stakeholders have come together to establish a new cross-faculty department for Digital Humanities at the University of Vienna. This activity is in line with the initiatives on the European level where a working group in DARIAH is developing a DH course registry and a reference curriculum for DH teaching and training.

3. ACDH-ÖAW

Rooted in the long tradition of RI activities for DH at the Academy (in particular at the ICLTT), there are plans to setup a whole new institute dedicated to this task – the Austrian Centre for Digital Humanities (based at the Austrian Academy of Sciences). This institute will grow out of the ICLTT's technical group, but is planned to be substantially expanded to better cover the whole range of digital humanities, especially archaeology and historical studies. The ACDH-ÖAW will assume the role of a national coordinator and represent Austria in international RI bodies.

[10] or ERIC – a new European legal entity for research infrastructures

[11] http://dhd2015.uni-graz.at/

The strategy is directed towards a tight interaction between ACDH-ÖAW and the other institutes of the Academy, bundling and sharing development resources and technical solutions (don't create a new repository for every institute or even project) in a matrix organization, i.e. staff from individual other institutes is also actively involved in the activities. Institutes delegate colleagues to cooperate with the ACDH on common solutions working on particular problems in projects running at the institutes.

4. Semantic turn

With the extraordinary diversity of disciplines and communities of practice that constitute the "digital humanities" a major challenge is to find a common language, a common understanding of the problems the various disciplines share. A promising technical approach to tackle the issue is the advancement of semantic technologies and the Linked Open Data paradigm (LOD, Berners-Lee, 2006). Although RDF and related technologies in themselves are not a universal remedy for all interoperability problems (rather just another form of information representation), it at least offers a common widely adopted syntactic denominator. Combined with the unifying force of the RIs on the organizational level this approach seems to have a high integrative, harmonizing potential.

In this respect, it has to be acknowledged that the bulk of existing research data exists in databases or XML-based formats, which means that before being able to take advantage of the new technology, a major effort is required to transform or enrich the data. This does not necessarily imply that all of the data needs to be converted into RDF right away. The change, the "semantic turn" as it is called by many, can and should happen gradually, in small steps.

As a first step, it may just be enough to semantically annotate existing data using well defined semantic reference resources as vocabularies, the trivial example being annotating persons as named entities in texts using the GND[12] (or dbpedia) resolvers. For data structured in databases it may be rather worthwhile to try to remodel the data in RDF, but here too, it needs to be decided if the added value is worth the effort. Adding a field with a URI identifying or classifying a given entity based on selected reference resources may be enough in the initial phase.

So, before moving into the world of complex ontologies, the existing data has to be normalized and enriched with links to semantic entities defined in well-established reference resources such as taxonomies or authority files. Accordingly, a number of activities have been started within CLARIN and DARIAH, both on the European and the national levels, aiming to coordinate the creation and maintenance of controlled vocabularies and other reference data. Within CLARIN, the initiative CLAVAS (Brugman, Lindeman, 2012) provides a number of vocabularies via a dedicated instance [13] of the open source vocabulary repository *OpenSKOS*[14] hosted by the Meertens Institute. Guided by the specific CLARIN needs, CLAVAS currently exposes the following vocabularies: a list of language codes (the ISO 639-3 standard converted to SKOS), a number of closed data categories taken from the data category registry *ISOcat*[15] and a list of organization names that has been extracted from the metadata collected from collaborating content providers. However, there exist multiple instances of *OpenSKOS* operated by different institutions offering a range of taxonomies, e.g. one managed by the Netherland Institute for Sound and Vision[16]. All of these are available via the same system and constitute a large pool of valuable reference data that can be tapped into at no cost, taking advantage of a uniform interface.

It is also planned to use the vocabulary repository *OpenSKOS* as a core module of the Knowledge Hub, a new integrative system for data and knowledge management that is currently being developed at the ACDH-ÖAW and will be available via CCV as an infrastructure service. In an integrated, largely automated environment metadata will be aggregated from a number of sources, it will be normalized and enriched using controlled vocabularies integrated in the system, before it is made available for browsing and searching (Ďurčo & Mörth, 2014).

As part of its CLARIN-DARIAH-AT commitment, the Austrian Audiovisual Research Archive (Austrian Academy of Sciences) has started to work on taxonomies (musical instruments, languages and language variants, geographical reference data). So far these have been used internally only and will be made publicly available in SKOS format. This data will be integrated into the ACDH instance of *OpenSKOS*. Another dataset to be included is the Taxonomy of Digital Research Activities in the Humanities or TaDiRAH[17] (Borek et al., 2014) which was developed in the DARIAH community. It is based on experiences in previous work in other projects like NeDiMAH and on Bamboo's DiRT taxonomy. It is already being used in the DH course registry [18] and in the bibliography collection on DH (*Doing digital humanities - a DARIAH bibliography* [19] *)*. The above mentioned vocabularies are only a starting point, the system will be open to new datasets, thus establishing an ever growing pool of reference data to be used internally and externally, both by applications and users. In adding new vocabularies, the focus lies on curation-intensive data such as for instance various named entities, e.g. organization names.

Another new service that has been launched recently as a DARIAH-DE contribution, offers a proxy service for the GND data (Gemeinsame Normdatei – the Integrated Authority File) which are maintained by the German National Library (GNL). The GND is a major normative reference resource in the German-speaking area and

[12] Gemeinsame Normdatei – the Integrated Authority File of the German National Library
[13] https://openskos.meertens.knaw.nl/
[14] http://openskos.org
[15] http://www.isocat.org
[16] http://openskos.beeldengeluid.nl/
[17] https://github.com/dhtaxonomy/TaDiRAH/
[18] http://dhcoursereg.hki.uni-koeln.de/
[19] https://www.zotero.org/groups/doing_digital_humanities_-_a_dariah_bibliography

beyond, however the native service endpoint provided by the GNL is available under very restrictive conditions. The unrestricted service which is made available by DARIAH-DE for the scientific community constitutes a major extension to the inventory of resources used for the task of semantic annotation.

The ACDH is involved in the activities as an early adopter and will be testing the new endpoint, planning to employ it in a number of ongoing DH projects requiring named entity recognition technology.

5. Conclusion and Outlook

As of 2014, a new phase in the institutional establishment of digital research infrastructures has begun. While on the European level CLARIN and DARIAH have both become official RI consortia formally installed by the European Commission, in Austria a new initiative was introduced to merge the hitherto rather fragmented activities, ensuring continuity by building on existing infrastructure components, but also breaking new ground through the orientation towards innovative cutting-edge technologies.

Next to the continuation of the "usual" RI work, ten new DH projects start this year which promises a substantial tide of new contributions in the years to come. The main challenge will be to safeguard the long-term preservation and availability of research data.

6. Acknowledgements

The initiative Austrian Centre for Digital Humanities / Digital Humanities Austria is funded and supported by the Federal Ministry of Science, Research and Economy and the partner institutions of CLARIN-DARIAH.AT[20].

7. References

AAC-Austrian Academy Corpus (2007). AAC-FACKEL Online Version: "Die Fackel. Herausgeber: Karl Kraus, Wien 1899-1936", http://www.aac.ac.at/fackel

Berners-Lee, T. (2006). Linked Data. online: http://www.w3.org/DesignIssues/LinkedData.html

Borek, L., Dombrowski, Q., Munson, M., Perkins, J. and Schöch, Ch. (2014). Scholarly primitives revisited: towards a practical taxonomy of digital humanities research activities and objects. In *Proceedings of Digital Humanities 2014*. Lausanne, Switzerland.

Broeder, D., Kemps-Snijders, M. et al. (2010). A data category registry- and component-based metadata framework. In M. Calzolari, N.; Choukri, K. & others (Eds.). *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010).* ELRA, Valetta.

Brugman, H. & Lindeman, M. (2012). Publishing and Exploiting Vocabularies using the OpenSKOS Repository Service. In *Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR Workshop Programme*, pp. 66.

Budin, G., Moerth, K. and Ďurčo, M. (2013). Working towards European Infrastructures - The ICLTT Language Resources Portal. In *49. Jahrestagung des Instituts für Deutsche Sprache*, Poster-Session, Korpora geschriebener Sprache. IDS, Mannheim.

Ďurčo, M. (2013). SMC4LRT - Semantic Mapping Component for Language Resources and Technology. Technical University, Vienna.

Ďurčo, M. & Windhouwer, M. (2014). From CLARIN Component Metadata to Linked Open Data. In *LDL 2014*, LREC Workshop. ELRA, Reykjavik.

Stehouwer, H., Ďurčo, M., Auer, E. and Broeder, D. (2012). Federated Search: Towards a Common Search Infrastructure. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*. pp. 3255-3259. ELRA, Istanbul.

Uytvanck, D. V., Zinn, C., Broeder, D., Wittenburg, P. and Gardellini (2010). Virtual Language Observatory: The Portal to the Language Resources and Technology Universe. In M. Calzolari, N.; Choukri, K. & others (Eds.). *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010).* ELRA, Valetta

[20] http://acdh.ac.at/consortium

Raziskovalna infrastruktura CLARIN.SI

Tomaž Erjavec[†], Jan Jona Javoršek[‡], Simon Krek[♠]

[†] Odsek za tehnologije znanja
[‡] Center za mrežno infrastrukturo
[♠] Laboratorij za umetno inteligenco
Institut »Jožef Stefan«
Jamova cesta 39, SI-1000 Ljubljana
tomaz.erjavec@ijs.si, jona.javorsek@ijs.si, simon.krek@ijs.si

Povzetek

V prispevku predstavimo slovensko jezikoslovno raziskovalno infrastrukturo CLARIN.SI, katere dolgoročni namen je, da v povezavi z evropsko infrastrukturo CLARIN ERIC spodbuja raziskave na področju humanistike in družboslovja s tem, da omogoči raziskovalcem enovit avtoriziran dostop do platforme, ki integrira jezikovne vire slovenskega jezika in napredna orodja za obdelavo slovenščine. Prispevek predstavi evropsko infrastrukturo CLARIN in njena temeljna načela ter povzame dosedanjo zgodovino vzpostavitve CLARIN.SI, nato pa podrobneje obdela trenutno stanje izgradnje te slovenske infrastrukture s poudarkom na repozitoriju jezikovnih virov in jezikoslovnih storitvah in orodjih.

The research infrastructure CLARIN.SI

The paper introduces the Slovene research infrastructure CLARIN.SI, whose long term objective is, in connection with the European research infrastructure CLARIN ERIC, to facilitate research in the humanities and social sciences by enabling researchers a uniform and authorised access to its platform, which will integrate Slovene language resources and advanced tools for processing of Slovene. The paper introduces CLARIN ERIC and its mission and summarises the history of the establishment of CLARIN.SI. It then discusses the current state of its development with a focus on the repository of language resources and on linguistic services and tools.

1. Uvod

CLARIN[1] (Váradi in dr., 2008) je ena izmed evropskih raziskovalnih infrastruktur, ki jih je izbral ESFRI, Evropski strateški forum o raziskovalnih infrastrukturah za Program evropskih raziskovalnih infrastruktur. CLARIN je distribuirana podatkovna infrastruktura, ki vključuje predvsem evropske univerze in raziskovalne inštitute. Od 2012 je CLARIN prijavljen kot evropska pravna oseba (CLARIN ERIC, European Research Infrastructure Consortium) in ima trenutno osem držav članic (Avstrija, Bolgarija, Češka, Nemčija, Danska, Estonija, Nizozemska in Poljska), deveta članica je meddržavno telo »Dutch Language Union«, članici pa bosta predvidoma kmalu postali tudi Norveška in Portugalska.

Kot piše na spletnih straneh CLARIN ERIC,[2] je dolgoročni namen te raziskovalne infrastrukture, da spodbuja raziskave na področju humanistike in družboslovja tako, da omogoči raziskovalcem enovit avtoriziran dostop do distribuirane platforme, ki integrira jezikovne vire in napredna orodja na evropski ravni. Ta vizija temelji na naslednjih stebrih:

1. ***Pokritje:*** V perspektivi naj bi vsak raziskovalec v humanistiki in družboslovju v EU in pridruženih članicah imel z enotnim overjanjem neposreden dostop do vseh zbirk digitalnih podatkov, ki vsebujejo na jeziku temelječa gradiva in so last oz. so dane v dostop s strani javnih ustanov.
2. ***Pravo:*** Za raziskave naj bi pri dostopu do podatkov ne bilo omejitev, razen tistih, ki izvirajo iz zaupnosti podatkov, pravice do zasebnosti ali etičnih zadržkov. Pravice in legitimni interesi lastnikov podatkov morajo biti zaščiteni.
3. ***Integracija podatkov:*** Iskanje po metapodatkih in vsebinah naj bi raziskovalcem omogočilo, da najdejo želene podatke. Lahko bodo gradili virtualne zbirke podatkov, ki prihajajo iz različnih virov in držav, in jih uporabljali, kot da bi bili vsi na istem mestu in z enakimi standardi zapisa.
4. ***Integracija storitev:*** Dodatno naj bi imeli raziskovalci tudi dostop do naprednih jezikovnotehnoloških storitev v obliki spletnih storitev, ki bi jim omogočili označevanje, raziskovanje, izkoriščanje, izboljšanje, upravljanje in vizualizacijo podatkov za podporo raziskovanju. Spletne storitve naj bi delovale na podatkih iz raznovrstnih virov, mogoče bi jih bilo sestavljati v kompleksne verige in strukture za izvedbo zahtevnih operacij.
5. ***Hramba:*** Rezultate raziskovalnih projektov in rezultate, dobljene z uporabo storitev, naj bi bilo možno shraniti kot nove podatkovne zbirke, tako da bi jih lahko uporabili tudi drugi raziskovalci. Podatki in rezultati naj bi bili trajnostno hranjeni in opremljeni s trajnimi identifikatorji, tako da bi bilo do njih možno dostopati za namen repliciranja rezultatov ali za izvajanje novih raziskav. Dodatno naj bi obstajale tudi trajne povezave do publikacij, ki uporabljajo ali dokumentirajo te vire.
6. ***Dostop:*** Raziskovalci naj bi razumeli in uporabljali infrastrukturo CLARIN brez tehničnih zadreg.
7. ***Brez meja:*** Infrastruktura CLARIN naj bi bila umeščena v globalno raziskovalno krajino in naj bi aktivno spodbujala preseganje meja med znanstvenimi področji, drugimi infrastrukturami, državami in kontinenti, kot tudi preseganje meja med akademskim in poslovnim svetom.

[1] http://www.clarin.eu

[2] http://www.clarin.eu/content/mission

Ta zelo ambiciozen načrt se je začel izvajati že leta 2008, do njegove uresničitve pa bo minilo še dosti časa.

V prispevku bomo predstavili, kako je z infrastrukturo CLARIN v Sloveniji, kjer se bomo navezali tudi na druga slovenska vozlišča evropskih humanističnih in družboslovnih raziskovalnih infrastruktur in na delo infrastrukture CLARIN v drugih evropskih državah. Prispevek v razdelku 2 obravnava zgodovino in ureditev infrastrukture CLARIN v Sloveniji, v razdelku 3 delo na vzpostavitvi repozitorija jezikovnih virov, v razdelku 4 spletne storitve, v razdelku 5 pa podamo nekaj zaključkov.

2. CLARIN v Sloveniji

Vlada RS je leta 2011 sprejela načrt razvoja slovenskih infrastruktur ESFRI, ki za humanistiko in družboslovje predvideva vzpostavitev slovenskih infrastruktur za DARIAH (Digital Research Infrastructure for the Arts and Humanities), ki je namenjena spodbujanju digitalno podprtih raziskav in poučevanja v humanističnih vedah in umetnosti, za CESSDA (Consortium of European Social Science Data Archives), ki opravlja podobno nalogo za družboslovje ter za CLARIN. Slovenski DARIAH in CESSDA sta bili ustanovljeni kmalu po sprejetju načrta, prejeli sta tudi financiranje in lahko pokažeta konkretne rezultate. Na Inštitutu za novejšo slovensko zgodovino (INZ) so v sodelovanju z Znanstvenoraziskovalnim centrom Slovenske akademije znanosti in umetnosti (ZRC SAZU) postavili spletno infrastrukturo SI-DIH,[3] ki omogoča iskanje podatkov po različnih repozitorijih oziroma arhivih institucij ali društev v humanistiki in umetnosti. Na Fakulteti za družbene vede Univerze v Ljubljani pa so vzpostavili spletno infrastrukturo ADP[4] (Arhiv družboslovnih podatkov), ki hrani zbirko podatkov, zanimivih za družboslovne analize, s poudarkom na problemih, povezanih s slovensko družbo.

Za razliko od DARIAH in CESSDA Slovenija ni bila vključena v prvo, pilotno fazo vzpostavljanja evropske infrastrukture CLARIN (2008–2011), zato je tudi realizacija vzpostavljanja infrastrukture v Sloveniji potekala zelo počasi, saj je minimalno financiranje steklo šele konec 2013 z Institutom »Jožef Stefan« (IJS) kot sedežem infrastrukture, pri čemer si upravljanje delita Odsek za tehnologije znanja E8 in Laboratorij za umetno inteligenco E3.

Začetek financiranja slovenske infrastrukture CLARIN je sovpadel s koncem velikega slovenskega projekta Sporazumevanje v slovenskem jeziku (SSJ), v okviru katerega je bilo v petih letih trajanja projekta zgrajenih večje število temeljnih jezikovnih virov in storitev za slovenski jezik (Arhar Holdt in dr., 2012; Krek in dr. 2012; Logar Berginc in dr., 2009). Spletišče projekta,[5] na katerem je možno uporabljati spletne storitve in prevzeti odprte jezikovne vire, je gostovalo pri podjetju Amebis, d.o.o., vendar je ob koncu projekta usahnilo financiranje za vzdrževanje strojne in programske opreme, kar je postavilo pod vprašaj nadaljnjo usodo dostopa do rezultatov projekta. Zato smo kot urgentno prvo nalogo infrastrukture postavili prenos spletišča na strežnike IJS, kar je vsebovalo nakup razmeroma zahtevne strojne in programske opreme, ki obsega 3 medsebojno povezane strežnike, od tega enega pod operacijskim sistemom Windows, dva pa GNU/Linux, ter prenos in namestitev programske opreme projekta. Čeprav navzven ni opaziti, razen hitrejšega delovanja, nobene razlike, je tako od začetka 2014 spletišče postavljeno na IJS, kjer se bo tudi naprej vzdrževalo.

V 2014 smo se tudi lotili vzpostavljanja formalnega statusa slovenske infrastrukture CLARIN, ki smo jo poimenovali CLARIN.SI. Sestanki na Ministrstvu za izobraževanje, znanost in šport Republike Slovenije ter s potencialnimi zainteresiranimi inštitucijami v Sloveniji so obrodili dobre rezultate: v začetku junija 2014 je devet partnerjev podpisalo Sporazum o ustanovitvi konzorcija CLARIN.SI. Konzorcij vključuje vse večje javne institucije kot tudi podjetja in društva, ki se ukvarjajo z jezikoslovjem in jezikovnimi tehnologijami v Sloveniji: Alpineon d. o. o.; Amebis, d. o. o.; Institut "Jožef Stefan"; Slovensko društvo za jezikovne tehnologije; Trojino, zavod za uporabno slovenistiko; Univerzo v Ljubljani; Univerzo v Mariboru; Univerzo na Primorskem in ZRC SAZU. S sporazumom je bil ustanovljen upravni odbor CLARIN.SI, v katerem ima vsaka članica enega zastopnika z namestniki in en glas pri glasovanju, s katerim se odloča o delovanju konzorcija. Upravni odbor je doslej imel en sestanek, na dopisnem glasovanju pa se je odločil, da med partnerje sprejme še INZ in Društvo za domače raziskave, snovalce in razvijalce spletnega slovarja Razvezani jezik.

Z vzpostavitvijo konzorcija je omogočeno, da Slovenija lahko zaprosi za včlanjenje v CLARIN ERIC in tako enakopravno sodeluje v delu evropskega konzorcija in izkorišča ugodnosti, ki jih nudi članstvo, npr. financiranje medsebojnih obiskov, sodelovanje v letnih srečanjih itd. Pogoj za vključitev je poleg zagotovitve tehničnih in pravnih pogojev tudi redno letno plačevanje članarine Slovenije, za katero je pristojno Ministrstvo za izobraževanje, znanost in šport.

3. Repozitorij jezikovnih virov

Eden od osnovnih storitev infrastrukture CLARIN je zagotavljanje zanesljivega arhiviranja in dostopa do jezikovnih virov, kot so korpusi, leksikoni, avdio in video posnetki, slovnice, jezikovni modeli itd. Za dolgoročno hranjenje jeziko(slo)vnih raziskovalnih podatkov so mnogi centri CLARIN po Evropi že vzpostavili storitve za deponiranje, ki vključujejo tudi pomoč pri tehničnih in organizacijskih zadregah, povezanih z deponiranjem. Storitve za deponiranje CLARIN naj bi imeli naslednje značilnosti:

1. *dolgoročno arhiviranje*: zagotovljen je dostop za daljše obdobje;
2. vire je mogoče enostavno citirati s *trajnimi identifikatorji*;
3. viri in njihovi metapodatki so integrirani v infrastrukturo, kar omogoča *učinkovito iskanje* po katalogih;
4. dostop do zaščitenih virov je omogočen preko *enotnega overjanja identitete uporabnikov*;
5. vire, integrirane v infrastrukturo CLARIN je možno analizirati in obogatiti z raznovrstnimi *jezikoslovnimi orodji*.

[3] http://www.sidih.si
[4] http://www.adp.fdv.uni-lj.si
[5] http://www.slovenscina.eu

Trenutno je v okviru CLARIN aktivnih dvanajst centrov za deponiranje in arhiviranje jezikovnih virov, pri čemer jih je sedem v Nemčiji, dva na Nizozemskem in po eden v Avstriji in na Češkem. Kljub enakim zunanjim tehničnim zahtevam so različni centri ubrali različne poti pri implementaciji arhivov, začenši z osnovno platformo za njihovo izgradnjo.

3.1. Repozitorij LINDAT

Za Slovenijo je bil najbolj zanimiv pristop, ki so ga ubrali na Češkem, kjer so v okviru Instituta za formalno in uporabno jezikoslovje Karlove univerze v Pragi (UFAL[6]) postavili servis LINDAT,[7] ki ima enostaven in uporabniku prijazen vmesnik in prinaša večino funkcij, ki jih želimo vključiti v sodoben repozitorij v okviru omrežja CLARIN. Za razvoj in vzdrževanje servisa LINDAT skrbi razmeroma velika ekipa, pri tem pa je češka različica tudi že pridobila »Data Seal of Approval«,[8] torej potrdilo, da izpolnjuje pogoje za trajen in zaupanja vreden digitalni repozitorij. LINDAT je odprtokodno dostopen in kolegi z UFAL so nam prijazno pomagali pri namestitvi repozitorija LINDAT na IJS.

LINDAT je osnovan na platformi za gradnjo digitalnih repozitorijev DSpace[9] (Branschofsky in dr., 2002), ki je odprtokodni projekt z velikim številom namestitev. DSpace je eden uspešnejših projektov za razvoj institucionalnih digitalnih repozitorijev, ki so v zadnjem desetletju in pol nastali kot odgovor na vse večje potrebe po organiziranem objavljanju, arhiviranju, bibliografski obdelavi in kuratorstvu digitalnih dokumentov v akademskem okolju. V raziskovalnem in akademskem okolju nove publikacije (članki in knjige) ne le nastajajo, temveč so vedno bolj pogosto tudi uporabljene ali vsaj distribuirane v elektronski obliki (Crow, 2002). DSpace temelji na konceptu »trajnih dokumentov« (durable document space) in se naslanja na priporočila referenčnega modela Open Archival Information Systems (OAIS, CCSDS 650.0-R-2, 2001) in priporočil FEDORA (2002), na osnovi katerih je nastal sistem Fedora Commons. Če primerjamo sistem DSpace s splošno sprejetimi zahtevami za takšne sisteme (Kenney in McGovern, 2003) ter s sorodnimi sistemi, zlasti odprtokodnim projektom Fedora Commons (prim. Lagoze in dr., 2006), ki smo ga že uporabili kot osnovni gradnik za postavitev repozitorija za digitalne dokumente, ter uveljavljenim sistemom GNU ePrints (Nixon, 2003; Kim, 2005), ima DSpace, še zlasti v prilagojeni različici LINDAT, kot repozitorij virov v okviru slovenskega vozlišča CLARIN nekaj očitnih prednosti.

Repozitorij omogoča ločeno obravnavo zahtev in avtorizacije več skupin uporabnikov. Vsak dokument je sestavljen iz metapodatkov v standardnem zapisu Dublin Core (Powell in Johnston, 2003) in enega ali več paketov (bundles), ki lahko vsebujejo enega ali več bitnih tokov (bit streams). DSpace datoteke shranjuje kot bitne tokove, paketi pa omogočajo združevanje datotek v logične skupine (npr. dokument v zapisu HTML s pripadajočimi slikami je logično en paket).

Za stabilen dostop do posameznega dokumenta oz. drugih deponiranih virov in njihovo navajanje je poskrbljeno z neodvisnim sistemom stabilnih identifikatorjev na osnovi sistema kazalcev (handles), ki ga razvija in vzdržuje Corporation for National Research Initiative (CNRI).[10] Sistem je povsem integriran z repozitorijem in poskrbi za dodeljevanje, upravljanje in razreševanje trajnih identifikatorjev za digitalne objekte in druge vire na internetu. Ker je sistem s katalogom kazalcev oz. identifikatorjev neodvisen od repozitorija, je torej mogoče v primeru spremembe domenskega sistema, zamenjave uporabljene arhitekture ipd. posodobiti naslove, kamor kažejo kazalci, in tako zagotoviti trajno veljavnost povezav na spletnih straneh in navedkov v objavljenih publikacijah. Uporaba takšnega sistema je pomembna zahteva za repozitorije v omrežju CLARIN, saj je mogoče na ta način zagotoviti trajno dostopnost virov in ponovljivost eksperimentov.

DSpace prinaša vrsto vtičnikov za registracijo in overjanje uporabnikov, ki preko mehanizma skupin omogočajo različne stopnje avtorizacije in različne vloge za uporabnike. Tako je mogoče določiti urednike ali administratorje posameznih zbirk, ki preko delotokov z uporabniki sodelujejo pri deponiranju virov in poskrbijo za uporabo ustreznih formatov in metapodatkov. V različici LINDAT je uporabljen prilagojen vtičnik, ki uporablja protokol SAML (Security Assertion Markup Language), ki se uporablja v sistemih za enotno spletno overjanje AAI (Authentication and Authorization Infrastructure). Vsaka organizacija tako lahko postane varuh osebnih podatkov svojih članov, ponudnikom aplikacij pa se ni treba ukvarjati z dodeljevanjem uporabniških imen ter kočljivim zbiranjem in preverjanjem podatkov o uporabnikih. Hkrati sistem omogoča posredovanje atributov uporabnika, tako da je mogoče članstvo v skupinah določati tudi na osnovi podatkov o uporabniku, ki jih posreduje njegova matična organizacija, npr. ARNES[11].

AAI je postal ena od ključnih tehnologij evropskih akademskih omrežij in skupnega evropskega raziskovalnega prostora, ker omogoča vzpostavitev nacionalnih (in širših) federacij, v katerih AAI povezuje uporabnike in storitve v celoto ter pridruženim organizacijam omogoča dodeljevanje enotnega uporabniškega imena, ki lahko uporabnikom služi za vse vrste aplikacij, tako v domači kot v drugih organizacijah v isti federaciji. Trenutni razvoj tehnologije (v okviru iniciative eduGAIN[12]) že omogoča tudi podporo za gostujoče uporabnike iz drugih nacionalnih federacij, (podobno kakor pri sorodni tehnologiji za overjanje v omrežju Eduroam), vendar ta sistem še ne deluje povsod. Zato vtičnik repozitorija LINDAT omogoča hkratno uporabo več kot ene identifikacijske federacije, kar je posebej pomembno, ker ima CLARIN lastno federacijo AAI, ki je nastala še pred iniciativo eduGAIN in tako omogoča dostop tudi uporabnikom, ki niso člani federacije AAI.

DSpace ima modularno arhitekturo za spletne vmesnike. LINDAT uporablja izvedbo spletnega vmesnika

[6] http://ufal.mff.cuni.cz
[7] https://lindat.mff.cuni.cz
[8] http://datasealofapproval.org
[9] http://www.dspace.org
[10] http://en.wikipedia.org/wiki/Handle_System
[11] https://aai.arnes.si
[12] http://www.geant.net/service/eduGAIN

XMLUI, ki uporablja podatkovne tokove XML in je zgrajena na osnovi odprtokodnega javanskega sistema za razvoj spletnih aplikacij Apache Cocoon.[13] Uveljavljena in fleksibilna tehnologija sicer dodaja nekaj kompleksnosti, vendar je razvoj repozitorija LINDAT dokaz, da je mogoče hitro in učinkovito razviti uporaben in uporabniku prijazen vmesnik.

Za CLARIN.SI trenutno poteka prilagajanje vmesnika s pomočjo mehanizmov, ki so jih razvijalci v okviru razvoja servisa LINDAT predvideli za prilagoditev na uporabo v drugih institucijah, ter lokalizacija vmesnika za uporabo v slovenščini. Zaradi narave implementacije vmesnika je ta naloga nekaj zahtevnejša, saj je treba vzporedno nekatere segmente lokalizirati v programskih paketih v jeziku Java (preko nastavitvenih datotek), druge pa je treba lokalizirati v podatkovnih tokovih XML v okviru spletnega vmesnika.

Zaradi hitrega razvoja na področjih podpore za overjanje AAI, prilagoditve na slovensko vozlišče CLARIN, lokalizacije, metodologije deponiranja virov in integracije novih razvojnih različic LINDAT je sicer pričakovati še občasne hitre in velike spremembe, vendar pa trenutna pilotna postavitev na osnovi razvojne različice izvedbe repozitorija LINDAT že deluje[14] in je primerna za testiranje in preizkušanje servisa. Do konca 2014predvidevamo prehod na stabilne različice in postopen začetek testne uporabe s pravimi podatki.

3.2. Deponiranje virov

Ko bo repozitorij CLARIN.SI postal operativen, bo treba zagotoviti, da bo ponujal dovolj kvalitetnih in za raziskovalce zanimivih jezikovnih virov. V prvi fazi nameravamo v repozitorij prenesti odprto dostopne vire projekta SSJ, torej take vire, ki jih je mogoče prevzeti (*download*) na lasten računalnik. Ti viri naj bi naknadno tudi služili kot primer dobre prakse, vključno s postopkom validacije prevzetih virov. Istočasno nameravamo v repozitorij vključiti odprtodostopne vire, ki jih na IJS trenutno ponujamo na spletnih straneh posameznih projektov, ki so omogočili njihov nastanek; primera sta ročno označena korpusa slovenskega jezika projekta JOS (Jezikoslovno označevanje slovenskega jezika)[15] (Erjavec in Krek, 2008) in korpusa ter besedišče starejše slovenščine projekta IMP[16] (Erjavec in dr., 2011). Na IJS imamo še več manjših (specializiranih, večjezičnih) korpusov in drugih jezikovnih virov, vendar pa se bo kmalu potrebno zazreti tudi navzven, v prvi meri h konzorcijskim partnerjem CLARIN.SI, ki so izdelali že večje število raznovrstnih virov slovenskega jezika, od korpusov do slovarjev.

Po sklepu upravnega odbora CLARIN.SI bodo do konca leta 2014 člani konzorcija pripravili seznam jezikovnih virov, ki bi jih bili pripravljeni prispevati v repozitorij skupaj z licenco oz. informacijo, kakšni so pogoji njihove uporabe.

Pri vključevanju teh virov v platformo pričakujemo dva problema. Prvi je tehnične narave, saj so jezikovni viri zapisani v raznovrstnih formatih, ki niso vedno dokumentirani. Rešitev vidimo v tehnični in finančni podpori bodisi za dokumentiranje zapisa virov (tu bomo morali biti posebej pozorni na dober izbor metapodatkovne sheme) in, kjer bo to le mogoče, konverzijo v enega od standardnih formatov, npr. TEI[17] (TEI, 2007) ali LMF[18] (Francopoulo, 2013).

Večji problem bodo verjetno predstavljale omejitve pri nadaljnjem razširjanju virov, ki so v lasti posameznih ustanov, kjer bodo problematične avtorske pravice in nenaklonjenosti ideji, da bi se viri hranili na javnem repozitoriju in bili na voljo kateremukoli raziskovalcu ali celo v komercialne namene. Tu bo vsaj v začetku verjetno potrebno vsak problem reševati posebej, nato pa se bodo sčasoma nabrale izkušnje in primeri dobrih praks, tako v Sloveniji kot tudi v ostalih nacionalnih repozitorijih CLARIN.

Pri virih, ki bodo nastali v prihodnje, bo situacija, upamo, bolj enostavna, vsaj če se bo sprejet (in se bo izvajal) predlog Akcijskega načrta za jezikovno opremljenost, ki predvideva korake za večjo odprtost izdelanih jezikovnih virov, ki nastanejo kot rezultat javnega financiranja.

Repozitorij CLARIN.SI bi lahko poleg jezikovnih virov vseboval tudi odprtokodne programe za jezikoslovne obdelave oz. modele zanje, ki podpirajo (tudi) slovenski jezik. Za razliko od jezikovnih virov verjetno ne bi bilo smiselno ponujati (le) prevzema programov, temveč tudi njihovo evidentiranje, skupaj s kazalko na sistem za upravljanje izvorne kode, kot sta GIT ali SVN, kjer poteka razvoj.

4. Jezikoslovna orodja in storitve

Poleg vzpostavitve repozitorija je naloga infrastrukture CLARIN tudi vzpostavitev sistema spletnih storitev, kar je (še) bolj dolgotrajen in kompleksen proces. Spletne storitve lahko razdelimo na take, ki so namenjeni vizualizaciji vsebine jezikovnih virov (konkordančniki za korpuse in pregledovalniki različnih vrst slovarjev oz. leksikalnih baz) in tiste, katerih namen je obdelati neki jezikovni vir, predvsem označiti korpus za nadaljnje analize.

4.1. Spletni dostop do korpusov

Na platformi CLARIN.SI predvidevamo povezave do obstoječih spletnih konkordančnikov za slovenščino in, kjer bo to mogoče, njihovo agregiranje. Tudi CLARIN.SI bo ponujal svoj konkordančnik oz. konkordančnike, kjer bomo kot tehnološko in korpusno osnovo vzeli že obstoječa konkordančnika nl.ijs.si, in sicer noSketchEngine[19] in CUWI[20], ki že sedaj ponujata preko 20 korpusov (Erjavec, 2013). Drugi dobro obiskani vmesniki do korpusov, ki bi jih tudi bilo smiselno vključiti v agregiran oz. enovit dostop, so v Sloveniji vsaj še:

- spletišče SSJ, z dostopom do korpusov Gigafida (reprezentativen), KRES (uravnotežen), Gos (govorni) in Šolar (učenci slovenščine s popravki napak);
- spletni vmesnik do Nove Beseda ZRC SAZU, ki ostaja zelo cenjen korpusni vir;

[13] http://cocoon.apache.org
[14] https://www.clarin.si/repository/xmlui/
[15] http://nl.ijs.si/jos
[16] http://nl.ijs.si/imp
[17] http://www.tei-c.org
[18] http://www.lexicalmarkupframework.org
[19] http://nl.ijs.si/noske
[20] http://nl.ijs.si/cuwi

- večjezični Evrokorpus Službe vlade RS za evropske zadeve, ki je povezan s terminološkimi slovarji.

Povezovanje zelo raznorodnih iskalnikov oz. korpusov je zanimiv in verjetno ne dokončno rešljiv problem, je pa v perspektivi zelo smiseln, saj bi uporabniku, ki bi rad izvedel nekaj o slovenskem jeziku, radi ponudili čim bolj reprezentativne podatke na enem mestu.

4.2. Slovarski in terminološki portali

Na institucijah, ki so članice infrastrukture CLARIN.SI, že obstajajo portali, ki ponujajo različne slovarske in terminološke vire. Taki primeri so:

1. viri Inštituta za slovenski jezik Frana Ramovša na bos.zrc-sazu.si:
   - Slovar slovenskega knjižnega jezika
   - Slovar novejšega besedja slovenskega jezika
   - Slovenski pravopis 2001
   - Pleteršnikov Slovensko-nemški slovar
   - Besede slovenskega jezika (združeno besedišče iz SSKJ, BSJ, korpusa Nova beseda in spletnega iskalnika NAJDI.SI)
   - Odzadnji slovar slovenskega jezika
   - Besedišče slovenskega jezika (BSJ - besede, ki niso bile sprejete v SSKJ)
2. slovarski viri na nl.ijs.si, izhajajoči iz različnih raziskovalnih projektov:
   - Japonsko-slovenski slovar za učence japonščine
   - Besedišče starejše slovenščine (iz ročno označenega korpusa starejše slovenščine IMP)
   - slovenski WordNet oz. sloWNet v spletni aplikaciji sloWTool
3. slovarski ali leksikonski viri SSJ na portalu www.slovenscina.eu:
   - leksikon besednih oblik Sloleks
   - leksikalna baza za slovenščino
4. portal Termania, www.termania.net podjetja Amebis, ki trenutno vsebuje 43 slovarjev, od terminoloških, dvojezičnih, splošnih itd.
5. Terminologišče, isjfr.zrc-sazu.si/terminologisce terminološke sekcije na Inštitutu za slovenski jezik Frana Ramovša, na katerem je mogoče dostopati do devetih terminoloških slovarjev.

Zaenkrat smo se odločili za prenos portala Termania na strežnike CLARIN.SI, saj kljub temu, da niti program niti vsebovani slovarji niso odprti, zagotavlja ta koristen in prostodostopen servis. V načrtu je tudi preverjanje možnosti agregiranega iskanja po zgoraj omenjenih portalih, s čimer bi bilo uporabnikom omogočeno poenoteno iskanje in skupen prikaz rezultatov. V daljši perspektivi bi pa bilo smiselno zasnovati splošni vmesnik do slovarskih podatkov, ki bi nato gostil čim večje število slovarjev.

4.3. Orodja in storitve za označevanje

Infrastrukture posameznih evropskih centrov že ponujajo dostop do orodij in spletnih storitev. Češki LINDAT ponuja za češčino (in druge jezike) poleg iskalnika po skladenjsko označenih korpusih (drevesnicah) tudi storitve za oblikoskladenjsko označevanje, označevalnik imenskih entitet, strojno prevajanje med češčino in slovaščino itd.

Tudi za slovenščino že obstaja nekaj orodij in spletnih storitev, npr. za oblikoskladenjsko označevanje in lematizacijo sta na voljo ToTaLe[21] projekta JOS in Obeliks[22], za skladenjsko analizo pa označevalnik SSJ[23]. V prihodnosti bi bilo dobro te in novo razvite označevalnike združiti v platformo CLARIN.SI in jih ponujati pod skupnim vmesnikom in omogočiti dostop do njih preko spletnih protokolov za izvajanje programov, kakršen je WSDL.

4.4. Spletni delotoki

Pri vedno večjem številu označevalnih in drugih spletnih storitev se kmalu pojavi potreba po njihovem dinamičnem kombiniranju; temu služijo platforme za izdelavo in izvajanje spletnih delotokov. Njihov razvoj je v zadnjem času postali zelo popularen, tudi na področju označevanja besedil. V okviru nacionalnih infrastruktur CLARIN so najdlje prišli v Nemčiji, kjer so na Univerzi v Tübingenu razvili sistem WebLicht[24], ki omogoča (predvsem za nemščino) izdelavo verige označevalnikov, pri kateri za posamezne korake lahko izbiramo med več sistemi.

Tudi na IJS že več let razvijamo platformo za delotoke ClowdFlows[25] (Kranjc in dr., 2012), ki je trenutno sicer usmerjena predvsem v podatkovno rudarjenje, v perspektivi pa bi lahko služila tudi kot podlaga za obdelavo besedil; pilotno smo sistem že preizkusili z orodjem ToTrTaLe (Pollak in dr., 2012).

Pri CLARIN.SI bi z implementacijo svoje platforme za izdelavo in izvajanje spletnih delotokov počakali na zadostno število virov in spletnih storitev, da ima izdelava sistema za njihovo dinamično kombiniranje smisel. Kompleksne operacije nad velikimi podatkovnimi množicami potrebujejo tudi velike računalniške kapacitete, kjer pa bi CLARIN.SI verjetno lahko uporabil slovenski nacionalni grid, SLING.[26]

Ker imajo v Nemčiji, kmalu pa mogoče tudi v drugih nacionalnih centrih CLARIN, takšne platforme z ustreznimi kapacitetami že na voljo, se tu pojavi tudi vprašanje, ali je smiselno vlagati v razvoj platforme CLARIN.SI za ustvarjanje in izvajanje delotokov, saj bi bilo verjetno bolj smiselno, da bi posamezne storitve, orodja ali modele za slovenščino enostavno ponudili v uporabo drugim platformam.

5. Zaključki

V prispevku smo predstavili prve korake slovenske raziskovalne infrastrukture CLARIN.SI[27] in načrte za nadaljnje delo po posameznih področjih, predvsem pri vzpostavljanju računalniške platforme in zagotavljanju virov in orodij, ki jo bodo osmislila.

V nadaljevanju pa bo seveda potrebno poskrbeti tudi za promoviranje CLARIN.SI, tako da se bodo tuji, predvsem pa domači raziskovalci, učitelji, študentje in drugi

[21] http://nl.ijs.si/analyse
[22] http://www.slovenscina.eu/tehnologije/oznacevalnik
[23] http://www.slovenscina.eu/tehnologije/razclenjevalnik
[24] http://weblicht.sfs.uni-tuebingen.de
[25] http://clowdflows.org
[26] http://www.sling.si
[27] http://www.clarin.si

potencialni uporabniki zavedali, da platforma obstaja in da jim lahko – upajmo – pomaga pri raziskovanju slovenskega jezika.

Zahvala

Avtorji se zahvaljujejo anonimnima recenzentoma za koristne pripombe. CLARIN.SI financira Ministrstvo za izobraževanje, znanost in šport Republike Slovenije.

Literatura

Branschofsky, M., Chudnov, D, 2002. "DSpace: Durable Digital Documents." V *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries* (New York: ACM), str. 372. doi:10.1145/544220.544319.
URL: http://hdl.handle.net/1721.1/26703

Crow, R., 2002. The Case for Institutional Repositories: A SPARC Position Paper. URL: http://www.sparc.arl.org/sites/default/files/media_files/instrepo.pdf (4. 7. 2014)

CCSDS 650.0-R-2: *Reference Model for an Open Archival Information System (OAIS)*. Red Book. Issue 2. June 2001.
URL: http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf.

Erjavec, T., Krek, S., 2008. Oblikoskladenjske specifikacije in označeni korpusi JOS. V T. Erjavec, J. Žganec Gros (ur.), *Zbornik šeste konference Jezikovne tehnologije*: Ljubljana: Institut Jožef Stefan, 49–53.

Erjavec, T., 2011. Automatic linguistic annotation of historical language: ToTrTaLe and XIX century Slovene. V zborniku: *5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Portland.

Erjavec, T., Jerele, I., Kodrič, M., 2011. Izdelava korpusa starejših slovenskih besedil v okviru projekta IMPACT. V: KRANJC, Simona (ur.). *Meddisciplinarnost v slovenistiki, (Obdobja, Simpozij, = Symposium, 30)*. Ljubljana: Znanstvena založba Filozofske fakultete, 41-47.

Erjavec, T., 2013. Korpusi in konkordančniki na strežniku nl.ijs.si. *Slovenščina 2.0*, 1 (1): 24–49. URL: http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_03.pdf

Flexible and Extensible Digital Object and Repository Architecture (FEDORA): URL: http://www.cs.cornell.edu/cdlrg/fedora.html

Francopoulo, G. (ur.), 2013. *LMF Lexical Markup Framework*. Wiley-ISTE.

Kenney, A. R., McGovern, N.Y., 2003. The Five Organizational Stages of Digital Preservation. V *Digital Libraries: A Vision for the 21st Century: A Festschrift in Honor of Wendy Lougee*, ur. Patricia Hodges, Mark Sandler, Maria Bonn, in John Price Wilkin, 122–53. Ann Arbor: The Scholarly Publishing Office, University of Michigan Library.

Kim, J., 2005. Finding Documents in a Digital Institutional Repository: DSpace and Eprints. V: *68th Annual Meeting of the American Society for Information Science and Technology (ASIST)*, Charlotte, ZDA, 28. 10. – 2. 11. 2005.

Kranjc, J., Podpečan, V., Lavrač, N., 2012. ClowdFlows: A Cloud Based Scientific Workflow Platform / Machine Learning and Knowledge Discovery in Databases. V *Lecture Notes in Computer Science*. Volume 7524. Springer, pp 816-819.

Krek, S., Grčar, M., Dobrovoljc, K., 2012. Označevalnik za slovenski jezik Obeliks. *Zbornik Osme konference Jezikovne tehnologije, 8. do 12. oktober 2012*, Ljubljana, Slovenia. 89-94.

Lagoze, C., Payette, S., Shin, E., Wilper C., 2006. "Fedora: An Architecture for Complex Objects and Their Relationships." International Journal on Digital Libraries 6(2): 124–38. doi:10.1007/s00799-005-0130-3.

Lewis, K. D., Lewis J. E., 2009. »Web Single Sign-On Authentication using SAML«. *IJCSI International Journal of Computer Science Issues*, zv. 2.

Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š., Krek, S., 2012. *Korpusi slovenskega jezika* Gigafida, KRES, ccGigafida *in* ccKRES*: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Nixon, W., 2003. DAEDALUS: Initial Experiences with Eprints and DSpace at the University of Glasgow. Ariadne, št. 37.
URL: http://www.ariadne.ac.uk/issue37/nixon (4. 7. 2014)

Pollak, S., Trdin, N., Vavpetič, A., Erjavec, T., 2012. NLP web services for Slovene and English: morphosyntactic tagging, lemmatisation and definition extraction. *Informatica*, 36/4, str. 441-449.

Powell, A., Johnston, P. 2003. *Guidelines for Implementing Dublin Core in XML*. URL: http://dublincore.org/documents/dc-xml-guidelines/ (4. 7. 2014)

Smith, MacKenzie, Bass, M., McClellan, G., Tansley, R., Barton, M., Branschofsky, M., Stuve, D., Harford Walker J., 2003. DSpace: An Open Source Dynamic Digital Repository, *D-Lib Magazine*, 9. zv., jan. 2003. URL: http://dlib.org/dlib/january03/smith/01smith.html (4. 7. 2014)

Tansley, R., Bass, M., Stuve, D., Branschofsky, M., Chudnov, D., McClellan, G., Smith, M., 2003. The DSpace institutional digital repository system: current functionality. *JCDL '03, Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, 87-97.

TEI Consortium (2007). TEI P5: Guidelines for Electronic Text Encoding and Interchange. URL: http://www.tei-c.org/Guidelines/P5.

Váradi, T., Krauwer, S., Wittenburg, P., Wynne, M., Koskenniemi, K. (2008). CLARIN: Common Language Resources and Technology Infrastructure. *6th International Conference on Language Resources and Evaluation (LREC 2008)*.

hrMWELex – a MWE lexicon of Croatian extracted from a parsed gigacorpus

Nikola Ljubešić[1], Kaja Dobrovoljc[2], Simon Krek[3], Marina Peršurić Antonić[4], Darja Fišer[4]

[1] Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences
University of Zagreb
I. Lučića 3, HR-10000 Zagreb
nljubesi@ffzg.hr

[2] Trojina, Institute for Applied Slovene Studies
Dunajska 116, SI-1000 Ljubljana
kaja.dobrovoljc@trojina.si

[3] Artificial Intelligence Laboratory
Jožef Stefan Institute
Jamova 39, SI-1000 Ljubljana
simon.krek@ijs.si

[4] Faculty of Arts
Aškerčeva 2, SI-1000 Ljubljana
mpersuric@gmail.com, darja.fiser@ff.uni-lj.si

Abstract

The paper presents the process of building the hrMWELex lexicon of multiword expressions extracted from the 1.9 billion-token parsed corpus of Croatian. The lexicon is built with the newly developed DepMWEx tool which uses dependency syntactic patterns to identify MWE candidates in parse trees. The extracted MWE candidates are subsequently scored by co-occurrence and organized by headwords producing a resource of more than 30 thousand headwords and 12 million MWE candidates. The evaluation of the lexicon showed an overall precision of just over 50% and quite varying precision over specific syntactic patterns. Finally, opportunities for the refinement and enrichment of this recall-high resource by distributional identification of non-transparent MWEs and cross-language linking are presented.

hrMWELex – Leksikon hrvaških večbesednih zvez, izluščenih iz skladenjsko označenega milijardnega korpusa

V prispevku predstavimo postopek izdelave leksikona hrMWELex, ki smo ga izluščili iz korpusa hrvaških besedil, ki je skladenjsko označen in vsebuje 1,9 milijarde besed. Leksikon smo zgradili s pomočjo orodja DepMWEx, ki za prepoznavanje kandidatov večbesednih zvez v odvisnostnih drevesih uporablja odvisnostne skladenjske vzorce, jih rangira in organizira glede na jedrno besedo. Izluščen leksikon vsebuje 30.000 jedrnih besed in 12 milijonov večbesednih zvez. Evalvacija leksikona pokaže natančnost luščenja, ki presega 50%, pri čemer natančnost pri različnih skladenjskih vzorcih zelo niha. Na koncu prispevka predstavimo možnosti za izboljšave in razširitev bogatega leksikona s pomočjo prepoznavanja netransparentnih večbesednih zvez s pomočjo načel distribucijske semantike ter možnosti povezovanja večbesednih zvez z ustreznicami v drugih jezikih.

1. Introduction

Multiword expressions (MWEs) are an important part of the lexicon of a language. There are various estimates on the number and therefore importance of MWEs in languages, but most claims point to the direction that the number of MWEs in a speaker's lexicon is of the same order of magnitude as the number of single words (Baldwin and Kim, 2010).

There are two basic approaches to identifying MWEs in corpora: the symbolic approach, which relies on describing MWEs through patterns on various grammatical levels, and the statistical approach, which relies on co-occurrence statistics (Sag et al., 2001). Most approaches take the middle road by defining filters through the symbolic approach and rank the candidates passing the symbolic filters by the statistical approach.

The two most frequently used grammatical levels used for describing MWEs are the one of morphosyntax and syntax (Baldwin and Kim, 2010). While morphosyntactic patterns (Church et al., 1991; Clear, 1993) are much more used since they have already yielded satisfactory results, there is a number of approaches that use the syntactic grammatical level as well (Seretan et al., 2003; Martens and Vandeghinste, 2010; Bejček et al., 2013).

In this paper we describe an approach that relies on syntactic patterns to identify MWE candidates. Our main argument for using the syntactic grammatical level is that on languages with partially free word order, such as Slavic languages, morphosyntactic patterns often have to rely on hacks, like allowing up to n non-content words between fixed words or classes, thereby keeping the precision under

control while at the same time trying not to loose too much recall. Still, a significant amount of recall is lost since often only the most frequent order of constituents of an MWE is taken into account.

On the other hand, an argument against using syntax for describing MWEs is the precision of the syntactic analysis which is around 80% for well-resourced Slavic languages while morphosyntactic description of well resourced Slavic languages regularly passes the 90% bar.

Most approaches that use the syntactic grammar layer for extracting MWEs, like (Pecina and Schlesinger, 2006) and the recently added feature in the well-known SketchEngine (Kilgarriff et al., 2004), take into account only MWEs consisting of two nodes, therefore missing the big opportunity syntax offers in defining much more complex patterns that could not be defined on the morphosyntactic level at all.

Until now, there were no efforts in producing large-scale MWE resources for Croatian. First experiments include (Tadić and Šojat, 2003) who use PoS filtering, lemmatization and mutual information to identify candidate terms as a preprocessing step for terminological work, (Delač et al., 2009) who experiment on a Croatian legislative corpus while developing the TermeX tool for collocation extraction and (Pinnis et al., 2012) who use the CollTerm tool, part of the ACCURAT toolkit, for extraction of terms as the first step in producing multilingual terminological resources. All the mentioned approaches use morphosyntactic patterns for identifying candidates and do not produce any resources. The only resource for Croatian that does rely on syntactic relations is the distributional memory DM.HR (Šnajder et al., 2013), whose primary goal is distributional modeling of meaning.

In this paper we describe our tool that enables writing complex dependency syntactic patterns for identifying MWE candidates and the resulting recall-oriented MWE resource obtained by applying the tool to a 1.9 billion-token parsed corpus of Croatian. As no such lexicon currently exists for Croatian and because it is unrealistic to expect heavy investment in similar resources in the near future, our goal is to build a universal resource that will be useful in a wide range of HLT (human language technologies) applications as well as to professional language service providers and the general public. We therefore aim to strike a balance between recall and precision, giving a slight preference to recall in the hope that, on the one hand, human users can deal with the errors efficiently, and applications on the other can resort to post-processing steps in order to mitigate negative effects of noise in the resource.

The paper is structured as follows: in the next section we describe the DepMWEx tool used in building the resource, in Section 3 we describe the resource in numbers and give its initial evaluation, in Section 4 we discuss further possibilities like calculating semantic transparency and taking a multilingual approach, and conclude the paper in Section 5.

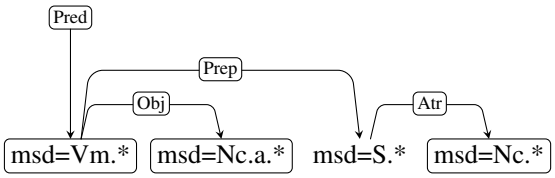


Figure 1: An example of the pattern tree corresponding to the MWE *tražiti rupu u zakonu*, *raditi račun bez konobara* (literally *to write the check without the waiter*), *raditi od buhe slona* (literally *make an elephant out of a fly*, *overexaggerate*) etc.

2. The DepMWEx tool

Our DepMWEx (Dependency Multiword Extractor) tool[1] consists of a Python module (defining the Tree and Node classes) and Python scripts that, given a grammar and a dependency parsed corpus, produce a list of strongest collocates for each headword.

2.1. The grammar

The grammar consists of a set of grammatical relations, each of which can be described with one or more so-called pattern trees.

Patterns trees are hierarchical structures in which each node contains a boolean function that defines the criterion a node in the parse tree of a sentence must satisfy to fill up that node. An example of a pattern tree, corresponding to the MWE *tražiti rupu u zakonu* (literally *search for a hole in the law*), which will be our working example in this section, is given in Figure 1. This pattern tree describes parse subtrees that have a predicate as a main verb which has direct object and prepositional phrase attached to it. The framed nodes represent headwords, i.e. for the example *tražiti rupu u zakonu*, this MWE candidate will be added to the headwords *tražiti#Vm*, *rupa#Nc* and *zakon#Nc*.

2.2. Grammatical relation naming

The name of the grammatical relation of our MWE example is "gbz sbz4 u sbz6", which is a notation taken over from the Slovene Sketch grammar (Kosem et al., 2013). That grammar is defined over morphosyntactic patterns, and, for reasons of compatibility, this Croatian grammar is based on that notation. The acronym denotes the part of speech ("gbz" being verb, "sbz" noun, "pbz" adjective and "rbz" adverb) while the number denotes the case, and "sbz4" stands for a noun in the accusative case. Finally, one can observe that in the grammatical relation the preposition is lexicalized, which is taken over from the Sketch grammar formalism.

Which part of the grammatical relation is the actual headword the MWE candidate occurs under is labeled by uppercasing that grammatical relation element, so under the verb *tražiti#Vm*, the MWE candidate *tražiti rupu u zakonu* will appear under the grammatical relation "GBZ sbz4 u sbz6".

[1] https://github.com/nljubesi/depmwex

2.3. Candidate extraction

The candidate extraction procedure is the following: over each parsed sentence from the corpus, each pattern tree makes an exhaustive search for sentence subtrees that satisfy its constraints. All subtrees corresponding to a pattern tree of a specific grammatical relation are written to standard output as (subtree, grammatical relation) pairs.

2.4. Candidate scoring

Once all (subtree, grammatical relation) pairs are extracted from the corpus, co-occurrence weighting is performed and MWE candidates are organized by their headwords and their grammatical relations. For now only the log-Dice measure (Rychlỳ, 2008), the association measure used in the Sketch Engine, is implemented in the tool. A selection of the resulting output for the headword *tražiti#Vm* is given in Table 1.

3. Resource description

3.1. The corpus

The lexicon was extracted from the second version of the Croatian Web corpus hrWaC (Ljubešić and Klubička, 2014), containing 1.9 billion tokens. The corpus was annotated with morphosyntactic, lemmatization and dependency parsing models built on the SETimes.HR manually annotated corpus (Agić and Ljubešić, 2014).

3.2. The grammar

The grammar for Croatian used in the DepMWEx tool was modified from the grammar for Slovene, which is based on the Slovene sketch grammar used in the SSJ project.[2] At this point the grammar consists of 63 grammatical relations defined through the same number of patterns trees. The constituents of the pattern trees are nouns in 53 relations, verbs in 33 relations, adjectives in 15 relations and adverbs in 11 relations.

3.3. The resulting lexicon

The resulting lexicon was filtered by the available lexical resources for Croatian, the Croatian morphological lexicon[3] and the Apertium morphological lexicon for Croatian.[4] Two frequency thresholds were enforced during the extraction process: the MWE candidate had to be of frequency 5 or higher, and the lexeme had to form at least 5 MWE candidates satisfying the first threshold. Entries for 46,293 lexemes (19,041 nouns, 11,183 adjectives, 7,028 verbs and 2,058 adverbs) were produced containing all together 12,750,029 MWE candidates. The relationship between the number of grammatical relations, the number of MWE candidates and the respective part of speech of the head is depicted in Figure 2. It shows that nouns are the most productive part of speech, being followed by verbs, adjectives and adverbs.

tražiti#Vm	logDice	freq
GBZ sbz4		
pomoć#Nc	8.358	9410
odšteta#Nc	7.958	1949
odgovor#Nc	7.851	4339
povrat#Nc	7.775	1952
ostavka#Nc	7.763	1900
zvijezda#Nc	7.503	2490
smjena#Nc	7.354	1385
rješenje#Nc	7.116	3127
posao#Nc	7.071	6353
naknada#Nc	7.031	1713
sbz1 GBZ sbz4		
prodavač#Nc način#Nc	8.457	330
tužiteljstvo#Nc kazna#Nc	7.295	147
čovjek#Nc mudrost#Nc	6.932	114
čovjek#Nc pomoć#Nc	6.840	108
sindikat#Nc povećanje#Nc	6.801	104
tužitelj#Nc kazna#Nc	6.575	89
prosvjednik#Nc ostavka#Nc	6.057	62
čovjek#Nc odgovor#Nc	6.001	60
žena#Nc muškarac#Nc	5.893	58
radnica#Nc pomoć#Nc	5.832	53
rbz GBZ		
uporno#Rg	7.589	715
stalno#Rg	7.579	1434
GBZ sbz4 za sbz4		
ponuda#Nc podizanje#Nc	10.831	587
rješenje#Nc problem#Nc	7.465	60
sredstvo#Nc ideja#Nc	6.995	39
stan#Nc najam#Nc	6.871	36
naknada#Nc šteta#Nc	6.869	36
obračun#Nc život#Nc	6.756	33
GBZ po sbz5		
vrlet#Nc	6.118	7
internet#Nc	5.612	227
džep#Nc	5.487	36
kontejner#Nc	5.334	29
oglasnik#Nc	4.718	10
kvart#Nc	4.714	21
inercija#Nc	4.623	5
forum#Nc	4.263	115
knjižara#Nc	4.181	8

Table 1: Part of the output of the DepMWEx tool for the headword tražiti#Vm

The final resource is encoded in XML and published[5] under the CC-BY-SA 3.0 license.

4. Initial resource evaluation

We performed an initial evaluation of the resource by inspecting up to 20 first MWE candidates for each grammatical relation of 12 selected lexemes. The analyzed lexemes were sampled as follows: 3 lexemes were taken for each part of speech, one in the upper, one in the medium and one

[2] http://eng.slovenscina.eu

[3] http://hml.ffzg.hr

[4] http://sourceforge.net/p/apertium/svn/HEAD/tree/languages/apertium-hbs/

[5] http://nlp.ffzg.hr/resources/lexicons/hrmwelex/

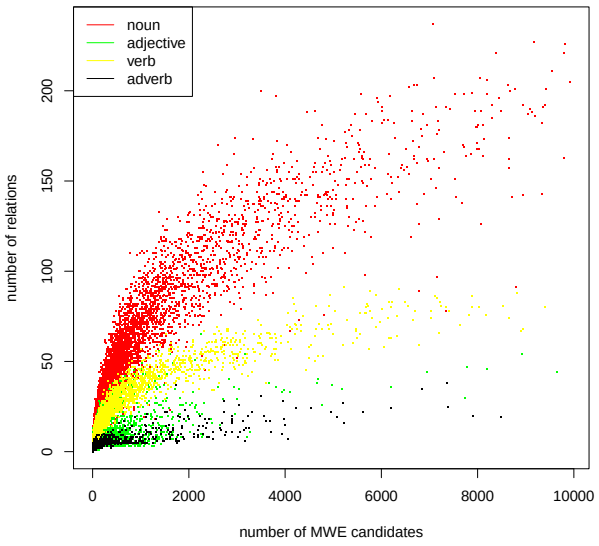


Figure 2: The relationship between the number of relations and MWE candidates by part-of-speech for each lexeme in the resulting lexicon

lexeme	# evaluated	precision
burza#Nc	559	0.735
lampa#Nc	154	0.422
lavež#Nc	34	0.324
N	747	0.652
gurati#Vm	311	0.296
razumjeti_se#Vm	161	0.484
tužiti_se#Vm	77	0.26
V	549	0.346
dužan#Ag	279	0.29
legendaran#Ag	64	0.609
svrhovit#Ag	20	0.4
A	363	0.353
naprosto#Rg	85	0.859
trostruko#Rg	78	0.615
jednoglasno#Rg	62	0.806
R	225	0.76
all	1884	0.518

Table 2: MWE candidate precision on each of the 12 evaluated lexemes

in the lower frequency range. We had one human annotator at our disposal annotating each MWE candidate as being a MWE or not. The precision obtained on each of the 12 lexemes, along with summaries for each part of speech and all lexemes, is given in Table 2. We can observe that the overall precision of the MWE candidates is just above 50% and that nouns and adverbs are more accurate than verbs and adjectives. Inside each part of speech the MWE candidate accuracies vary significantly and there is no correlation between the frequency range of a lexeme and its precision (the lexemes are ordered by falling frequency).

Next, we analyzed the precision of each specific grammatical relation. The precision for each grammatical relation occurring 10 or more times in the 12 lexemes is given in Table 3. The worst performing set of grammatical relations are the "in/ali" (*and/or*) relations which search for the same-POS constituents combined with the *and* or *or* conjunction. Another frequent and poorly performing relation is the one of a noun subject and its main verb predicate when the verb is the head (sbz1 GBZ) while significantly better results (0.64 vs. 0.167) are obtained with the subject as the head of relation (SBZ1 gbz). A similar phenomenon can be observed with the grammatical relation consisting of a main verb and its direct object which is performing very poorly when the verb is considered the head of the relation (GBZ sbz4), but with noun as head (gbz SBZ4), the obtained precision is much higher (0.214 vs. 0.714). This result stresses the fact that some relations are actually not symmetric and that the relations as they are defined now have to be reconsidered in the future.

5. Lexicon refinement

At this point we produced a recall-high resource with satisfactory precision, just over 50%, and the next obvious step is additional filtering of the resource with the goal of getting the precision rate up without hurting recall. Besides filtering, classifying the MWE candidates into types of MWEs should be looked into as well.

5.1. Semantic transparency

One of the properties of MWEs we are especially interested in is semantic transparency. We have already performed initial experiments in identifying that type of idiosyncrasy by using the distributional approach.

We built context vectors for all MWE candidates that fall under the following grammatical relations: "pbz0 SBZ0", "SBZ0 sbz2" and "VBZ sbz4". Besides building context vectors for MWE candidates, we also built vectors for their heads.

We built context vectors from three content words to the left and right, stopping at sentence boundaries. We took into consideration only MWE candidates occurring 50 times or more, which we consider minimum context information for any prediction. We used TF-IDF for weighting the vector features and Dice similarity for comparing vectors. We obtained the IDF statistic from head context vectors. The full procedure applied in calculating semantic transparency is the following:

1. build the frequency context vector for each MWE and its head
2. subtract the MWE vector frequencies from the headword vector (thereby remove contextual information of that MWE)
3. transform both vectors to TF-IDF vectors
4. calculate the Dice similarity score between each MWE and its head

By inspecting MWE candidates, organized under their heads and ordered by the computed similarity to the head, we observed quite promising results. We give a few examples for the simplest "pbz0 SBZ0" relation:

relation	frequency	precision
pbz0 SBZ0	94	0.809
RBZ gbz	73	0.822
RBZ pbz0	65	0.923
rbz GBZ	60	0.5
sbz1 GBZ	60	0.167
RBZ RBZ	52	0.558
SBZ1 gbz	50	0.64
GBZ u sbz5	49	0.204
GBZ0 in/ali GBZ0	47	0.213
PBZ0 in/ali PBZ0	47	0.277
GBZ na sbz4	46	0.283
SBZ0 in/ali SBZ0	45	0.0
gbz SBZ4	42	0.714
GBZ sbz4	42	0.214
rbz PBZ0	42	0.357
sbz0 SBZ2	42	0.667
GBZ u sbz4	41	0.829
SBZ0 sbz2	32	0.656
RBZ Vez-gbz pbz1	27	0.704
gbz Inf-GBZ	25	0.64
SBZ0 u sbz5	24	0.208
gbz na SBZ4	23	0.652
gbz na SBZ5	22	0.727
rbz Vez-gbz PBZ1	22	0.227
SBZ1 gbz sbz4	22	0.864
sbz0 na SBZ5	20	0.9
PBZ0 Inf-gbz	20	0.85
gbz s SBZ2	20	1.0
sbz0 na SBZ4	20	0.7
sbz0 s SBZ2	20	0.95
PBZ0 u sbz5	20	0.05
gbz sbz4 na SBZ5	20	0.85
pbz0 na SBZ5	20	1.0
GBZ sbz6	19	0.421
PBZ0 za sbz4	18	0.278
SBZ0 na sbz5	17	0.765
SBZ0 za sbz4	17	0.529
SBZ0 od sbz2	16	0.375
PBZ0 sbz6	16	0.125
PBZ0 prije sbz2	15	0.6
GBZ sbz4 u sbz4	14	0.5
PBZ0 na sbz4	13	0.154
PBZ0 po sbz5	13	0.308
SBZ0 s sbz6	13	0.615
GBZ do sbz2	12	0.417
SBZ0 o sbz5	12	1.0
PBZ0 na sbz5	12	0.083
PBZ0 o sbz5	11	0.182
sbz0 za SBZ4	11	0.818
GBZ prema sbz5	11	0.455
sbz1 gbz SBZ4	10	0.9
SBZ0 u sbz4	10	0.8
sbz1 GBZ sbz4	10	0.3
gbz preko SBZ2	10	1.0
GBZ s sbz6	10	0.6
PBZ0 od sbz2	10	0.1

Table 3: Precision scores per grammatical relations (sorted by frequency)

- for the head *voda* (water), the most distant MWE candidate is *amaterska voda* (*amaterske vode* refers to a person who moves from professional to amateur), the second one being *Baška voda* (a municipality in Croatia)
- for the head *selo* (village), the two most distant MWE candidates are *Novo Selo Žumberačko* (a municipality) and *špansko selo* (refers to something absolutely unknown to someone, like *it's all Greek to me*)
- for the head *stan* (flat) the least similar MWEs are *vječni stan* (*eternal resting place*, an experimental dark music album and the Catholic metaphor for heaven), *Ninski stanovi* (a municipality) and *tkalački stan* (sewing machine)
- for the head *ured* (office), the most distant MWE is *ovalni ured* (the Oval office)
- for the head *sastanak* (meeting), the most distant MWE is *Brijunski sastanak* (an important meeting during the Croatian independence war)
- for the head *zlato* (gold), among the most distant MWEs are *tekuće zlato* (referring to any liquid which is very valuable) and *crno zlato* (referring to oil)

On the other hand, once we sorted all the results, regardless of their head, the results seem much less usable. Besides non-transparent MWEs, we obtain probable parsing errors, low-frequency entries, entries with very static context etc. Nevertheless, the obtained results can be very useful for a lexicographer inspecting a specific headword and will therefore be added to the new version of the lexicon.

5.2. Multilinguality

We have already made first inquiries in the multilingual setting by producing similar lexicons for two other south Slavic languages, namely Slovene [6] and Serbian[7], but using smaller amounts of data. Since the grammatical relations have the same names in grammars of all the languages, we can use *(grammatical relation, dependents)* pairs as features for our context vectors, obtaining therefore a more detailed and selective formalization of the context of a lexeme than in the standard distributional approach as implemented in the previous subsection. We thereby possibly form more potent distributional memories (Baroni and Lenci, 2010) for tasks of inducing multilingual lexicons of closely related languages by using lexical overlap or similarity, as was done in (Ljubešić and Fišer, 2011). It would be interesting to inspect how such a memory compares to the already existing distributional memory of Croatian DM.HR (Šnajder et al., 2013) which takes into account only binary relations.

We give here one example for the Croatian–Serbian language pair. The Serbian noun *vaspitanje* is not present

[6] http://nlp.ffzg.hr/resources/lexicons/slmwelex/

[7] http://nlp.ffzg.hr/resources/lexicons/srmwelex/

in Croatian, but by observing its strongest MWE candidates, which are for the relation "sbz0 SBZ2" *nastava*, *profesor*, *nastavnik* and for the relation "pbz0 SBZ0" *fizički*, *predškolski*, *građanski*, for a human it becomes obvious that the two Croatian counterparts are *odgoj* and *obrazovanje*, which have very similar entries under the same grammatical relations, such as *uvođenje*, *nastava* and *nastavnik* for the "sbz0 SBZ2" relation and *predškolski*, *zdravstven* and *građanski* for the "pbz0 SBZ0" relation. If a model was constructed by using *(grammatical relation, dependent)* pairs as features and log-Dice as their weights, the models of those two lexemes on the Croatian side would have an overwhelming similarity with the Serbian lexeme in comparison to other lexeme combinations with that Serbian lexeme.

6. Conclusion

In this paper we presented the process of building a recall-oriented MWE lexicon of Croatian with the newly developed DepMWELex tool which uses syntactic patterns for MWE candidate extraction. Although MWEs are an important part of a lexicon of a certain language, and often key for proficient knowledge and use of a language, they are still not sufficiently represented in dictionaries, lexicons and other resources. This is especially the case with Croatian and other under-resourced languages. Thus the intention of building this MWE lexicon was to build a MWE resource that has a wide range of use, including HLT applications, professionals and the general public. Such an extensive resource offers a vast array of possibilities of researching the Croatian language and its MWEs. Learners of Croatian, as well as professional translators translating into Croatian as their non-mother tongue lack such a resource.

Since the recall-high approach was taken in producing the resource, the overall precision of the candidates lies slightly above 50%. Nevertheless, there are big differences in accuracies of specific grammatical relation, so a lexicon with precision of $\sim$ 80% can be produced easily by just filtering out the noisy grammatical relations.

The possibility of calculating semantic transparency of MWE candidates with the distributional approach is inspected as well with very promising results on the lexeme level. Using the produced output for modeling the context of a lexeme and using it for cross-language linking is shown off as well.

This work presents just the first step towards a rich MWE resource of not just Croatian, but its neighboring languages as well. Future work on the resource will start with increasing the size of the underlying corpora for the lexicons of Slovene and Serbian and publishing a three-language resource. For that resource to be of maximum value, the possibilities of cross-language linking on both the headword and MWE candidate levels with the distributional approach will be looked into. Finally, focused research on identifying non-transparent MWEs will be undertaken as well.

7. References

Željko Agić and Nikola Ljubešić. 2014. The SETimes.HR linguistically annotated corpus of Croatian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Eduard Bejček, Pavel Stranak, and Pavel Pecina. 2013. Syntactic identification of occurrences of multiword expressions in text using a lexicon with dependency structures. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 106–115, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. 1991. Using statistics in lexical analysis. In *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 115–164. Erlbaum.

Jeremy Clear, 1993. *Text and Technology: In honour of John Sinclair*, chapter From Firth Principles - Computational Tools for the Study of Collocation. John Benjamins Publishing Company.

Davor Delač, Zoran Krleža, Jan Šnajder, Bojana Dalbelo Bašić, and Frane Šarić. 2009. Termex: A tool for collocation extraction. In Alexander F. Gelbukh, editor, *CICLing*, volume 5449 of *Lecture Notes in Computer Science*, pages 149–157. Springer.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. *Information Technology*, 105:116.

Iztok Kosem, Simon Krek, and Polona Gantar. 2013. Automatic extraction of data: Slovenian case revisited. In *SKEW-4: 4th International Sketch Engine Workshop*, Talinn, Estonia.

Nikola Ljubešić and Darja Fišer. 2011. Bootstrapping bilingual lexicons from comparable corpora for closely related languages. In *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, volume 6836 of *Lecture Notes in Computer Science*, pages 91–98. Springer.

Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden. Association for Computational Linguistics.

Scott Martens and Vincent Vandeghinste. 2010. An efficient, generic approach to extracting multi-word expressions from dependency trees. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 85–88, Beijing, China, August. Coling 2010 Organizing Committee.

Pavel Pecina and Pavel Schlesinger. 2006. Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL on Main Conference*

Poster Sessions, COLING-ACL '06, pages 651–658. Association for Computational Linguistics.

Mārcis Pinnis, Nikola Ljubešić, Dan Ştefănescu, Inguna Skadiņa, Marko Tadić, and Tatiana Gornostay. 2012. Term extraction, tagging, and mapping tools for under-resourced languages. In *Proceedings of the Terminology and Knowledge Engineering (TKE2012) Conference*, Madrid, Spain.

Pavel Rychlỳ. 2008. A lexicographer-friendly association score. *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, pages 6–9.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multiword expressions: A pain in the neck for nlp. In *In Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002*, pages 1–15.

Violeta Seretan, Luka Nerima, and Eric Wehrli. 2003. Extraction of multi-word collocations using syntactic bigram composition. In *In Proceedings of the International Conference RANLP'03*, pages 424–431.

Jan Šnajder, Sebastian Padó, and Željko Agić. 2013. Building and evaluating a distributional memory for croatian. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789. Association for Computational Linguistics.

Marko Tadić and Krešimir Šojat. 2003. Finding multiword term candidates in croatian. In *Proceedings of Information Extraction for Slavic Languages 2003 Workshop*, pages 102–107.

Determining the Semantic Compositionality of Croatian Multiword Expressions

Petra Almić and Jan Šnajder

University of Zagreb, Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
petra.almic@gmail.com, jan.snajder@fer.hr

Abstract

A distinguishing feature of many multiword expressions (MWEs) is their semantic non-compositionality. Being able to automatically determine the semantic (non-)compositionality of MWEs is important for many natural language processing tasks. We address the task of determining the semantic compositionality of Croatian MWEs. We adopt a composition-based approach within the distributional semantics framework. We build a small dataset of Croatian MWE with human-annotated semantic compositionality scores. We build and evaluate a model for predicting the semantic compositionality based on Latent Semantic Analysis. The predicted scores correlate well with human judgments (ρ=0.48). When compositionality detection is treated as a classification task, the model achieves an F1-score of 0.65.

Določanje semantične kompozicionalnosti hrvaških večbesednih enot

Pomembna lastnost številnih večbesednih enot je njihova semantična nekompozicionalnost. Zmožnost avtomatskega določevanja takšne (ne)kompozicionalnosti je pomembna za številne naloge pri obdelavi naravnega jezika. V prispevku obravnavamo določanje semantične kompozicionalnosti hrvaških večbesednih enot. Uporabimo metodo, ki temelji na kompozicionalnosti v okviru distribucijske semantike. Zgradimo majhno podatkovno množico hrvaških večbesednih enot z ročno določenimi vrednostmi njihove semantične kompozicionalnosti. Zgradimo in evalviramo model za napovedovanje semantične kompozicionalnosti, ki temelji na latentni semantični analizi. Napovedane vrednosti dobro korelirajo s človeškimi ocenami (ρ = 0,48). Če detektiranje kompozicionalnosti obravnavamo kot klasifikacijsko nalogo, doseže model za mero F1 vrednost 0,65.

1. Introduction

The peculiarity of multiword expressions (MWEs) has long been acknowledged in natural language processing (NLP). According to Sag et al. (2002), MWEs can be defined as idiosyncratic interpretations that cross word boundaries (or spaces). Because of their unpredictable and idiosyncratic behavior, such expressions need to be listed in a lexicon and treated as a single unit ("word with spaces") (Evert, 2008; Baldwin et al., 2003). One dimension along which the MWEs can be analyzed is their semantic compositionality, sometimes referred to as semantic idiomaticity or semantic transparency. Semantic compositionality is the degree to which the features of the parts of an MWE combine to predict the features of the whole (Baldwin, 2006). The meaning of a non-compositional MWE cannot be deduced from the meaning of its parts. In reality, MWEs span a continuum between completely compositional expressions (e.g., *world war*) to non-compositional ones (Bannard et al., 2003). A prime example of non-compositional MWEs are idioms, such as *kick the bucket (to die)* or *red tape* (excessive rules and regulations).

Being able to determine the semantic compositionality of MWEs has been shown to be important for many NLP tasks, ranging from machine translation (Carpuat and Diab, 2010) and information retrieval (Acosta et al., 2011) to word sense disambiguation (Finlayson and Kulkarni, 2011). It is thus not surprising that the task of automatically determining semantic compositionality has gained a lot of attention (Katz and Giesbrecht, 2006; Baldwin, 2006; Biemann and Giesbrecht, 2011; Reddy et al., 2011; Krčmář et al., 2013).

In this paper we address the task of automatically determining the semantic compositionality of Croatian MWEs comprised of two words. We follow up on the work of Katz and Giesbrecht (2006) and Biemann and Giesbrecht (2011) and adopt a compositionality-based approach. The basic idea is to compare the meaning of an MWE against the meaning of the composition of its parts. To model the meaning of the MWEs and its parts, we use distributional semantics, which represents the word's meaning based on the distribution on its contexts in a corpus, assuming that similar words tend to appear in similar contexts (Harris, 1954). To determine the compositionality of an MWE, we compare its context distribution in a corpus to the context distribution approximated by the composition of its parts.

The contribution of our work is twofold. Firstly, we build a dataset of Croatian MWE annotated with semantic compositionality scores. Second, we build and evaluate a semantic compositionality model based on Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997). Our results show that the compositionality scores produced by the model correlate well with human-annotated scores, thereby confirming similar results for the English language. To the best of our knowledge, this is the first work to consider semantic compositionality detection for the Croatian language.

2. Related work

The approaches to determining the semantic compositionality can be broadly divided into two groups: knowledge-based approaches and corpus-based approaches. The former rely on linguistic resources (e.g., WordNet) to measure the semantic similarity between an MWE and its parts (Kim and

Baldwin, 2006). The obvious downside of knowledge-based approaches is that the linguistic resources are unavailable for the most languages and that acquiring them is expensive. In contrast, corpus-based approaches rely on statistical properties of MWEs and the constituting words, which can be readily extracted from corpora. E.g., McCarthy et al. (2007) rely on the hypothesis that non-compositional MWEs tend to be syntactically more fixed than compositional MWEs, while Pedersen (2011) assumes that lexical association correlates with non-compositionality.

Related to the work presented in this paper are the corpus-based approaches that rely on the distributional semantic modeling of MWEs and their constituents. The pioneering work in this direction is that of Lin (1999), who used a statistical association measure to discriminate between compositional and non-compositional MWEs. Lin compared the mutual information of an MWE and of an expression obtained as a slight modification of the original MWE (e.g., *red tape* vs. *orange tape*). Although this method has not shown to be successful, the idea that non-compositional expressions have a "different distributional characteristic" than similar compositional expressions paved a way for other distributional semantics based approaches. Baldwin et al. (2003) used LSA to compare the similarity between an MWE and its head, and showed that there exists a correlation between the measured semantic similarity and compositionality. Along the same lines, Katz and Giesbrecht (2006) used LSA to compare the semantic vector of an MWE against the semantic vector of the composition of its constituents, obtained simply as the sum of the corresponding vectors.

To consolidate the research efforts, Biemann and Giesbrecht (2011) organized a shared task on semantic compositionality detection, and provided datasets in English and German with human compositionality judgments. The task was shown to be hard and no clear winner emerged. However, the approaches based on distributional semantics seemed to outperform those based on statistical association measures. Shortly thereafter, Krčmář et al. (2013) performed a systematic evaluation of various distributional semantic approaches to compositionality detection, and showed that LSA-based models perform quite well.

In this paper we adopt the methodology of Katz and Giesbrecht (2006) to compare the distribution of an MWE to the composition of its parts, but we experiment with different composition functions, proposed by Mitchell and Lapata (2010). To build the dataset, we adopt the methodology of Biemann and Giesbrecht (2011).

3. Annotated dataset

The starting point of our work is a dataset of representative Croatian MWEs annotated with human compositionality judgments. In building this dataset, we adopted the approach of Biemann and Giesbrecht (2011), but depart from it in some key aspects that we discuss below. As a source of data, we used the 1.2 billion words corpus fHrWaC[1] (Šnajder et al., 2013), a filtered version of the Croatian web corpus *hrWaC* (Ljubešić and Erjavec, 2011). The corpus has been tokenized, lemmatized, POS tagged, and dependency parsed using the the HunPos tagger and the CST lemmatizer for Croatian (Agić et al., 2013), and the MSTParser for Croatian (Agić and Merkler, 2013), respectively. We next describe the construction of the dataset.[2]

3.1. MWE extraction

Following the work of Biemann and Giesbrecht (2011), we restricted ourselves to the following three MWE types:

- **AN**: an adjective modifying a noun, e.g., *žuti karton* (*yellow card*);
- **SV**: a verb with a noun in the subject position, e.g., *podatak govori* (*data says*);
- **VO**: a verb with a noun in the object position, e.g., *popiti kavu* (*drink coffee*).

We extracted all dependency bigrams (i.e., possibly non-contiguous bigrams) from the corpus that match one of these three types and sorted them by frequency in descending order.[3] Going from the top of list, we (the two authors) manually annotated the MWEs and additionally pre-annotated each as compositional (C) or non-compositional (NC). We next selected the bigrams on which both annotators agreed, and then balanced the set so that it contains an equal number of compositional and non-compositional MWEs. The so-obtained dataset does not reflect the true distribution of MWEs, as the compositional MWEs are much more frequent in the corpus. However, as our focus is on discriminating between the compositional and non-compositional MWEs, balancing the dataset is justified in this case. The final dataset contains 100 compositional and 100 non-compositional MWEs (125 AN, 10 SV, and 65 VO expressions). Note that the C/NC annotation is preliminary; each of the 200 MWEs has subsequently been annotated with compositionality scores by multiple human annotators other than the authors (cf. Section 3.3.).

3.2. Levels of compositionality

During the process of the candidate selection, we identified various flavors of compositionality. For example, a *yellow card* really is a yellow card, but it has an additional (and a dominant one) figurative meaning (a warning indication). In contrast, *gray economy* is indeed a type of economy, but *gray* does not stand for a color here. Further along these lines, *chain* in a *chain store* is not a chain in its dominant sense. One can argue that all these expressions are non-compositional to a certain extent. In an attempt to give an operational account of the different levels of non-compositionality, we propose the following typology:[4]

[1] `http://takelab.fer.hr/data/fhrwac/`

[2] The dataset is available under the Creative Commons BY-SA license from `http://takelab.fer.hr/cromwesc`

[3] By considering only the most frequent MWEs, we limit ourselves to MWEs with most reliable distributional representations.

[4] Note that our typology is motivated by practical rather than theoretical concerns. In the realm of automatic compositionality detection, type NC3 is arguably more easily determinable than type NC1. From a theoretical perspective, the proposed typology is oversimplified and we make no attempts here to relate it to the different types of figures of speech studied in linguistics (e.g., metaphors, metonyms, synegdochs, etc.).

NC3: Expressions that are completely non-compositional, i.e., the meaning of constituents cannot be combined to give the meaning of the expression. E.g., *žuti karton (yellow card)* and *preliti čašu* (literal meaning: *spill over the cup*; figurative meaning: *the last straw*), *trljati ruke (to rub ones hands)*;

NC2: Partially compositional expressions, i.e., the meaning of one but not both constituents is opaque, e.g., *siva ekonomija (gray economy)*, *bilježiti rast (to record a growth)*, *morski pas* (literal meaning: *sea dog*; compositional meaning: *a shark*);

NC1: The expressions that are non-compositional if we consider only the dominant senses of one or both of its constituents. For example, if we consider a *chain* only as a series of metal rings, then a *mountain chain* is a non-compositional expression.[5]

We (the two authors) annotated the 200 MWEs according to the above types and resolved the disagreements by consensus. Our primarily motivation for this was to be able to investigate how the level of non-compositionality influences the performance of the model.

3.3. Annotation

Biemann and Giesbrecht (2011) used the crowdsourcing service Amazon Turk to annotate their dataset. For every expression, they provided five different context sentences. For each in-context MWE, they asked the turkers to annotate how literal the MWE is, on a scale from 1 (non-compositional) to 10 (compositional). Because of this setup, they were not able to estimate the inter-annotator agreement, but they argued that the judgments for the expressions should be reliable because they were averaged over several sentences and several annotators. As the final compositionality scores, they computed the mean score for each MWE.

We departed from the above-described setup for two reasons. Methodologically, we argue that annotating MWEs across contexts is inappropriate for the task of semantic compositionality detection of the sort we are addressing here. The reason is that it ignores the fact that MWEs may have different meanings (compositional and non-compositional ones) depending on the context, thus averaging across the contexts will lump together the various senses. On a practical side, in-context annotation is more expensive and would require more resources (we feel that annotating five sentences per MWE would not suffice to reliably capture the sense variability of MWEs). For these reasons, we chose not to annotate MWEs across different contexts.

Our annotation setup was as follows. A total of 24 volunteers (mostly students) participated in the annotation. To reduce the workload, we divided the 200 MWEs into four groups (A, B, C, D) and randomly assigned one group to each annotator. Thus, each MWE was annotated by six annotators. To be able to computer the inter-annotator agreement, we ensured a 10% overlap among all four groups (20 expressions that were annotated by all 24 annotators).

[5]We are aware that the notion of a dominant sense is a problematic one. Many of the NC1 MWEs in our dataset are in fact borderline cases between NC and C classes.

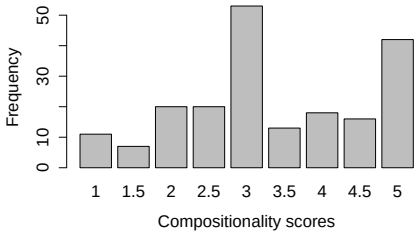


Figure 1: Histogram of MWE compositionality scores.

MWE	Score
maslinovo ulje (olive oil)	5
krvni tlak (blood pressure)	5
telefonska linije (telephone line)	4
pružiti pomoć (to offer help)	4
kućni ljubimac (a pet)	3.5
crno tržište (black market)	3
voditi brigu (to worry)	3
ostaviti dojam (to leave an impression)	2.5
zeleno svjetlo (green light)	1
hladni rat (cold war)	1

Table 1: Examples from the annotated dataset.

We asked our annotators to judge how literal each MWE is on the scale from 1 (non-compositional) to 5 (compositional). For each MWE, we provided one context sentence that instantiates its non-compositional meaning (for non-compositional MWEs) or typical compositional meaning (for compositional MWEs). We did this to ensure that annotators consider the same sense of an MWE, so that the judgments would not diverge because of sense mismatches.

We computed the final compositionality score for each MWE as the median of its compositionality scores. Fig. 1 shows the scores histogram, while Table 1 shows some examples from the annotated dataset.

3.4. Annotation analysis

Table 2 shows the inter-annotator agreement in terms of the Krippendorff's alpha coefficient (Krippendorff, 2004) for each of the groups as well as the overlapping part of the dataset. We consider the agreement to be moderate and indicative of the high subjectivity of the task. The agreement on the verb expressions is somewhat lower in comparison to adjective-noun expressions. In Table 3 we present some example MWEs from the dataset where the annotators achieved a high level of agreement (zero standard deviation) and a low level of agreement (st. dev. > 1.3).

Sample	AN+SV+VO	AN	SV+VO
Group A	0.587	0.620	0.535
Group B	0.506	0.510	0.478
Group C	0.490	0.544	0.337
Group D	0.586	0.505	0.648
Overlap (10%)	0.456	0.452	0.439

Table 2: Inter-annotator agreement (Krippendorff's α).

High agreement	Low agreement
igrati nogomet (play soccer)	*zabilježiti rast (record growth)*
služiti kaznu (serve sentence)	*žuti karton (yellow card)*
financijska pomoć (financ. aid)	*prvi korak (first step)*
pjevati pjesmu (sing song)	*telefonska linija (phone line)*
nemati sumnje (have no doubt)	*crveni karton (red card)*

Table 3: Examples of MWEs with high and low inter-annotator agreement on compositionality scores.

To be able to compare the performance of the models against human judgments as the ceiling performance, we computed the correlation between every annotator's scores and the median scores. The average Spearman's correlation coefficient over 24 annotators is 0.77.

4. Compositionality model

To build our model, we use the fHrWaC corpus, the same corpus we used to build the dataset. To optimize and experiment with the various parameters, we randomly split our dataset into the train and test set, each consisting of 100 MWEs. To determine the semantic compositionality of a MWE, we carry out the following three steps: (1) model the meaning of the constituent words, (2) model the composition of the meaning, and (3) compare these meanings.

Modeling word meaning. To model the meaning of constituent words, we use the Latent Semantic Analysis (Landauer and Dumais, 1997). LSA has shown to perform quite good in the task of semantic compositionality detection (Katz and Giesbrecht, 2006; Krčmář et al., 2013). Furthermore, LSA models excelled in the task of identifying synonyms in the Croatian language (Karan et al., 2012). We defined the context as a ± 5 word window around the word, or, in the case of the MWEs, a ± 5 word window around both constituents. For the constituent words, we only considered the contexts in which they appear alone, i.e., not as a part of any MWE from our dataset. Motivation behind this is to emphasize the independent contribution of the constituents in an expression, as proposed by Katz and Giesbrecht (2006). As context elements (the columns of the LSA matrix), we use the 10k most frequent lemmas from the corpus (excluding stop words). As target elements (the rows of the matrix), we used the MWEs and their constituting words, as well as the 5k most frequent lemmas from the corpus. For weighing the word-context associations, we experimented with two functions: log-entropy (Landauer, 2007) and Local Mutual Information (LMI) (Evert, 2005). We used singular value decomposition to reduce the dimensionality of the matrix from 10000 to 100 dimensions per target.

Modeling composed meaning. The second step was to model the composition of the word meanings. Mitchell and Lapata (2010) introduced a number of composition models (additive, weighted additive, multiplicative, tensor product, and dilation), which they evaluated on a phrase similarity task (e.g. *vast amount* vs. *large quantity*). In this work, we experiment with additive ($\vec{z} = \vec{x} + \vec{y}$), weighted additive ($\vec{z} = \alpha\vec{x} + \beta\vec{y}$), and the multiplicative model ($\vec{z} = \vec{x} \odot \vec{y}$), where z stands for the composed vector and $\vec{x}$ and $\vec{y}$ stand for vectors of its constituent words.

We experiment with two weighted additive models. In the first one (model Opt), similarly to Mitchell and Lapata (2010), we optimized the weights on the train set to maximize the correlation with human scores. The weights are optimized globally and they are identical for every MWE. In the second one (model Dyn), we calculated the weights dynamically, separately for each MWE, as proposed by Reddy et al. (2011). The two weights, α and β, are defined as

$$\alpha = \frac{\cos(\overrightarrow{xy}, \vec{x})}{\cos(\overrightarrow{xy}, \vec{x}) + \cos(\overrightarrow{xy}, \vec{y})}, \quad \beta = 1 - \alpha \qquad (1)$$

where $\overrightarrow{xy}$ is the MWE vector. The intuition behind this method is that more importance should be given to the constituent that is semantically more similar to the whole MWE, i.e., the constituent whose vector is closer, in terms of the cosine similarity, to the vector of the MWE. For example, in the expression *gray economy*, more importance should given to the word *economy* than the word *gray*.

In addition, we experiment with a linear combination of the additive model, the multiplicative model, and the two individual constituents model (Reddy et al., 2011):

$$\begin{aligned}\lambda = a_0 &+ a_1 \cdot \cos(\overrightarrow{xy}, \overrightarrow{x+y}) + a_2 \cdot \cos(\overrightarrow{xy}, \overrightarrow{x \odot y}) \\ &+ a_3 \cdot \cos(\overrightarrow{xy}, \overrightarrow{x}) + a_4 \cdot \cos(\overrightarrow{xy}, \overrightarrow{y})\end{aligned} \qquad (2)$$

We optimized the parameters a_0–a_4 using least squares regression on the train set.

Meaning comparison. Finally, in the third step, we use the cosine similarity measure to compare the vector-represented meaning of the MWE and the vector of its composition-derived meaning. We expected that for the compositional MWEs these two meaning vectors will be similar, i.e., cosine similarity will be closer to 1, while for non-compositional it will be closer to 0.

5. Evaluation

The task of determining semantic compositionality can be framed as a regression problem (prediction of compositionality scores) or a classification problem (compositionality vs. non-compositionality). We consider both settings.

5.1. Predicting compositionality scores

In Table 4 we show the correlation (Spearman's ρ) between model-predicted and human-annotated compositionality scores on the test set. Even though we experimented with two weighting functions, here we present only the results for log-entropy because LMI gave consistently worse results. Additive models outperform the multiplicative model. This is in contrast to the conclusions of Mitchell and Lapata (2010), but in accordance with the results of Guevara (2011) and Krčmář et al. (2013). Also, it is noticeable that the AN expressions have better correlation than verb expressions, which goes along the fact that the former had a higher inter-annotator agreement. Best performing model is the linear combination, which suggests that combining the evidence from multiple models is beneficial. Overall, results seem to be comparable to the results in (Biemann and Giesbrecht, 2011; Krčmář et al., 2013) obtained for English.

Model	AN+SV+VO	AN	SV+VO
Multiplicative	−0.19	−0.20	−0.18
Simple additive	0.45	0.54	0.35
Weighted additive (Opt)	0.46	0.56	0.28
Weighted additive (Dyn)	0.46	0.57	0.26
First constituent	0.41	0.50	0.19
Second constituent	0.28	0.31	0.31
Linear combination (λ)	**0.48**	**0.56**	**0.34**
Annotators	0.77	0.77	0.74

Table 4: Correlation results on the test set.

	AN+SV+VO	AN	SV+VO
Precision	0.58	0.74	0.43
Recall	0.73	0.65	0.77
Accuracy	0.65	0.72	0.54
F1-score	0.65	0.69	0.56

Table 5: Classification results on the test set.

5.2. Compositionality classification

For the compositionality classification task, we converted the compositionality scores to binary labels. To this end, we analyzed the distribution of the scores in the dataset (Fig. 1). Because the distribution is bimodal, we decided to set the cut-off after the first peak, so that MWEs with the score in the $[1, 3]$ range are labeled as non-compositional (NC), while those with the score in the $\langle 3, 5]$ range are labeled as compositional (C). We consider only the best-performing model from the previous evaluation task (the Linear combination model). The model predicts C if the cosine similarity between the MWE vector and the linear combination vector is above a certain threshold, otherwise it predicts NC. We optimized the threshold on the train set by optimizing the F1-score. The results are shown in Table 5.

The classification task is similar to the one considered by Katz and Giesbrecht (2006). In their experiment, they achieved the F1-score of 0.48, but they only considered the additive model for modeling semantic compositionality.

5.3. Result analysis

In this section we give some insights about the model performance. Results show moderate level of correlation, so we are interested in investigating on what MWEs the model fails. We are also interested in relating the model performance to the levels of compositionality introduced in Section 3.2. and the inter-annotator agreement levels.

In Table 6 we list the MWEs on which the model performs the worst. We define the error as an absolute difference in the Z-scores between the model-predicted and human-annotated scores. The results seem to suggest that most errors occur on compositional expressions (C), which happen to be the ones on which the annotators easily agreed about the high degree of compositionality.

To explore this hypothesis a bit further, we divided our test set into the subsets based on the compositionality levels (C – 48%, NC1 – 31%, NC2 – 7%, NC3 – 14%), and then calculated correlation on each subset separately. Fig. 2 shows the (absolute) correlation on each of these subsets, for high and low inter-annotator agreement levels. The plot again suggests that the model performs the worst on the compositional MWEs, while it performs best on partially non-compositional MWEs.

MWE	Prediction	Error	Level
nemati sumnje	2.48	2.85	C
organizacijski odbor	2.66	2.56	C
dati život	2.16	2.55	NC3
optužnica teretiti	4.51	2.51	C
spasiti život	2.85	2.25	C
uroditi plodom	3.85	2.24	NC1
izvršna vlast	2.61	2.23	C

Table 6: MWEs on which the model performs the worst.

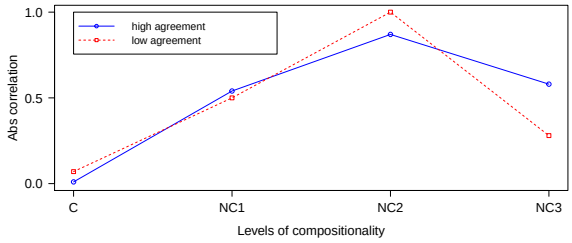


Figure 2: Correlation on the test set for the four compositionality levels and two inter-annotator agreement levels.

A deeper analysis should be done to determine the underlying causes. One of the possible reasons could be the low quality of vector representations for some (rare) words. The low quality of the individual words propagates to the low quality of compositional representations, which in turn makes the composed vector too dissimilar to the MWE vector. A further problem might stem from the polysemy, another weakness of distributional semantic models.

6. Conclusion

We considered the problem of determining the semantic compositionality of Croatian multiword expressions (MWEs) using a composition-based distributional semantics approach. We built a small dataset of Croatian MWEs, manually annotated with semantic compositionality scores. To represent the meaning of the MWEs and their constituents, we built an LSA model over the Croatian web corpus. We experimented with the additive and multiplicative compositional models. The best-performing model combines the additive and the multiplicative compositional models and the representations of the two individual words. The model achieves a correlation of 0.48 and an F1-score of 0.65.

For future work we plan to enlarge the dataset to allow for a more reliable analysis. Furthermore, we will consider doing the analysis on an unbalanced and hence a more realistic dataset. We also intend to consider the task of token-based semantic compositionality detection, along the lines of Cook et al. (2007) and Sporleder and Li (2009).

7. References

O. C. Acosta, A. Villavicencio, and V. P. Moreira. 2011. Identification and treatment of multiword expressions applied to information retrieval. In *Proc. of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, pages 101–109. ACL.

Ž. Agić and D. Merkler. 2013. Three syntactic formalisms for data-driven dependency parsing of Croatian. In *Text, Speech, and Dialogue*, pages 560–567. Springer.

Ž. Agić, N. Ljubešić, and D. Merkler. 2013. Lemmatization and morphosyntactic tagging of Croatian and Serbian. In *Proc. of ACL*.

T. Baldwin, C. Bannard, T. Tanaka, and D. Widdows. 2003. An empirical model of multiword expression decomposability. In *Proc. of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 89–96. ACL.

T. Baldwin. 2006. Compositionality and multiword expressions: Six of one, half a dozen of the other. In *Invited talk given at the COLING/ACL'06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*.

C. Bannard, T. Baldwin, and A. Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proc. of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18*, MWE '03, pages 65–72. ACL.

C. Biemann and E. Giesbrecht. 2011. Distributional semantics and compositionality 2011: Shared task description and results. In *Proc. of the Workshop on Distributional Semantics and Compositionality*, pages 21–28. ACL.

M. Carpuat and M. Diab. 2010. Task-based evaluation of multiword expressions: A pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 242–245. ACL.

P. Cook, A. Fazly, and S. Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proc. of the Workshop on a Broader Perspective on Multiword Expressions*, pages 41–48. ACL.

S. Evert. 2005. *The statistics of word cooccurrences*. Ph.D. thesis, Dissertation, Stuttgart University.

S. Evert. 2008. Corpora and collocations. *Corpus Linguistics. An International Handbook*, 2:223–233.

M. A. Finlayson and N. Kulkarni. 2011. Detecting multiword expressions improves word sense disambiguation. In *Proc. of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, pages 20–24.

E. Guevara. 2011. Computing semantic compositionality in distributional semantics. In *Proc. of the Ninth International Conference on Computational Semantics*, pages 135–144. ACL.

Z. Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

M. Karan, J. Šnajder, and B. D. Bašić. 2012. Distributional semantics approach to detecting synonyms in Croatian language. *Information Society*, pages 111–116.

G. Katz and E. Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proc. of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19. ACL.

S. N. Kim and T. Baldwin. 2006. Automatic identification of English verb particle constructions using linguistic features. In *Proc. of the Third ACL-SIGSEM Workshop on Prepositions*, pages 65–72. ACL.

K. Krippendorff. 2004. Reliability in content analysis. *Human Communication Research*, 30(3):411–433.

L. Krčmář, K. Ježek, and P. Pecina. 2013. Determining compositionality of expresssions using various word space models and methods. In *Proc. of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 64–73. ACL.

T. K. Landauer and S. T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, pages 211–240.

Landauer. 2007. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates.

D. Lin. 1999. Automatic identification of non-compositional phrases. In *Proc. of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 317–324. ACL.

N. Ljubešić and T. Erjavec. 2011. hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. In *Text, Speech and Dialogue*, pages 395–402. Springer.

D. McCarthy, S. Venkatapathy, and A. K. Joshi. 2007. Detecting compositionality of verb-object combinations using selectional preferences. In *EMNLP-CoNLL*, pages 369–379.

J. Mitchell and M. Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

T. Pedersen. 2011. Identifying collocations to measure compositionality: shared task system description. In *Proc. of the Workshop on Distributional Semantics and Compositionality*, pages 33–37. ACL.

S. Reddy, D. McCarthy, S. Manandhar, and S. Gella. 2011. Exemplar-based word-space model for compositionality detection: Shared task system description. In *Proc. of the Workshop on Distributional Semantics and Compositionality*, pages 54–60. ACL.

I. A. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer.

J. Šnajder, S. Padó, and Ž. Agić. 2013. Building and evaluating a distributional memory for Croatian. In *In Proc. of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 784–789. ACL.

C. Sporleder and L. Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proc. of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 754–762. ACL.

Approximate Measures in the Culinary Domain: Ontology and Lexical Resources

Cvetana Krstev,* Staša Vujičić Stanković,† Duško Vitas†,

* Faculty of Philology, University of Belgrade
Studentski trg 3, 11000 Belgrade, Serbia
cvetana@matf.bg.ac.rs
† Faculty of Mathematics, University of Belgrade
Studentski trg 16, 11000 Belgrade, Serbia
(stasa,vitas)@matf.bg.ac.rs

Abstract

Language resource development is extremely important for Serbian, as a less-resourced language, to take it into the digital era. In our research we focused on the culinary domain, given the increasing popularity of linguistic processing of culinary content. We provide a detailed description of the language resources – electronic morphological dictionaries, the WordNet semantic network, and a corpus of Serbian written culinary recipes, developed during our earlier work, as well as our latest efforts in enriching morphological dictionaries and WordNet with approximate measure terminology and developing an approximate measure ontology. The paper presents the issues related to detecting and categorizing the approximate measures from the culinary domain to be marked with new domain-specific semantic markers and populate the ontology, and indicates the benefits that language resources gain after addressing them.

Približne mere v kulinariki: ontologija in leksikalni viri

Za srbščino kot jezik s pomanjkljivo jezikovno opremljenostjo je razvoj jezikovnih virov izrednega pomena, saj bo le tako uspešno prešla v digitalno dobo. V naši raziskavi smo se osredotočili na področje kulinarike, saj je mogoče zaslediti vedno večje zanimanje za jezikoslovno obdelavo besedil s tega področja. V prispevku podamo natančen opis jezikovnih virov – računalniških morfoloških slovarjev, semantičnega leksikona WordNet in korpusa srbskih kuharskih receptov, ki so bili razviti v predhodnem delu, kot tudi naše trenutne raziskave o razširitvi morfoloških slovarjev in leksikona WordNet s terminologijo približnih mer in izgradnjo ontologije približnih mer. Prispevek predstavi problematiko identifikacije in kategorizacije približnih mer iz domene kulinarike, ki jih je treba označiti z novimi, domensko specifičnimi semantičnimi markerji in vključiti v ontologijo, ter pokaže na prednosti za jezikovne vire, ki jih prinese razreševanje te problematike.

1. Introduction

It seems that the culinary domain is one of the rare domains in which the general public and the scientific community are equally interested today. The first claim can be easily supported by a number of web sites, which offer a huge number of recipes, in many languages, searchable by different criteria, and often populated by users. A number of such sites exist in Serbian as well. Moreover, many TV shows worldwide are devoted to the art of cooking. In addition to popular magazines, the publishing of culinary books and manuals is still flourishing: from at least 70 to more than 200 such works are published each year in Serbia and recorded by the National Library of Serbia.

On the other hand, Various aspects of the culinary domain continuously attract the research community. The existence of various scientific institutions[1] and many scientific publications[2] from the domain can serve as evidence. The new application from IBM "Chef Watson with Bon Appétit" uses Watson's capabilities to explore big data to create new recipes.[3]

It is obvious that such an attractive and vivid domain is interesting for information processing (Mori et al., 2012; Dufour-Lussier et al., 2012; Wiegand et al., 2012; Ahnert, 2013; Nedovic, 2013), as it offers a lot of resources in the form of written and spoken texts. Obviously, it also has to be supported by information technologies, like ontologies (Cantais et al., 2005; Batista et al., 2012; Kim, 2012), in order to build various applications (for example systems based on ontologies like FOODS (Snae and Bruckner, 2008), TAAABLE (Badra et al., 2008) or Global Track&Trace Information System (Pizzuti et al., 2014)).

In this paper, we will first present Serbian language resources that are not only used for the processing of texts from the culinary domain, but also benefit from it (Section 2). Next, we will present one specific aspect of this domain, namely the use of measures, in recipes (Section 3) with special emphasis on the approximate, more informal, measures that are not listed in formal standards or professional manuals (e.g. 'a pinch of', 'small bunch', 'clove of' etc.) (Section 4). Section 5 presents an approximate measures ontology and gives the details about how they are covered in the Serbian resources. Finally, we will show how an adequate treatment of measures helps in the processing of texts from the culinary domain and give some directions for the future work (Section 6).

2. Serbian Language Resources in the Culinary Domain

2.1. The Corpus of Serbian Written Culinary Recipes

For the purpose of research into the culinary domain, we created a corpus of approximately 14,000 culinary recipes (more than 1.5 million word forms) in Serbian (both pronunciations – Ekavian and Ijekavian), written in

[1] One of the most important is IEHCA – Institut européen d'histoire et des cultures de l'alimentation, in Tours, France.
[2] The IEHCA catologue can be consulted at http://www.portail.scd.univ-tours.fr and the selected scientific bibliography at http://www.foodbibliography.eu.
[3] See http://www.research.ibm.com/software/IBMResearch/multimedia/Cognitive-Cooking-Fact-Sheet.pdf.

the Latin script. The recipes were drawn from *Recepti*[4] and other similar Serbian culinary Internet portals mentioned above.

As any web user interested in food preparation can post her/his recipes on these sites, their content, regarding both their style and syntax, is not strictly controlled. Therefore different types of errors were identified. The most frequent one that cannot be automatically corrected, at least not in all cases, is the omission of Latin script diacritics or their replacement with digraphs,[5] which introduces a number of homographic forms. To resolve this problem, we did not include in our corpus any recipe that does not feature at least one Serbian Latin script letter with diacritics. In the remaining recipes, we managed to recover some of the missing diacritics with the help of Serbian e-dictionaries (described in the next subsection).

2.2. Serbian Electronic Dictionaries

The basic resources for natural language processing of Serbian consisting of electronic (e-)dictionaries and local grammars are being developed using the finite-state methodology as described in (Courtois et al., 1990). The main role of these resources is text tagging. Each word form in an e-dictionary is equipped with the following information: (a) lemma; (b) Part-of-Speech (PoS); (c) set of values of grammatical categories pertinent to a PoS; (d) set of markers – syntactic, semantic, dialectic, derivational, domain etc. – describing a lemma. As reported in (Krstev, 2008), the system of Serbian electronic dictionaries covers both general lexica and proper names, and its present version is derived from 131,000 simple form lemmas and 13,000 compound lemmas (a.k.a. multi word units). In addition to that, a collection of finite-state transducers (FSTs) has been developed to support tagging that recognizes multi-word units belonging to open sets, e.g. multi word numerals and other numerical expressions (Krstev & Vitas, 2006).

Semantic marker	**Description**
+Culinary	culinary domain
+Food	food
+Alim	alimentation (e.g. *žito* ‘wheat’)
+Prod	product (e.g. *brašno* ‘flour’)
+Course	course (e.g. *torta* ‘cake’)
+Ing	ingredient (e.g. *so* ‘salt’)
+Meal	meal (e.g. *čajanka* ‘tea party’)
+Uten	utensil (e.g. *ekspres-lonac* ‘express pot’)
+Taste	taste (e.g. *aromatizovan* ‘flavored’)
+WoP	way of preparation (e.g. *nadevati* ‘stuff’ and *nadeven* ‘stuffed’)
+Cond	condition (e.g. *taze* ‘fresh’)
+MesApp	approximate measure (e.g. *kriška* ‘slice’)

Table 1. The semantic markers in Serbian e-dictionaries related to the culinary domain.

In order to improve the processing of texts from the culinary domain, we enlarged our e-dictionaries with new lemmas from this domain and systematically added the appropriate semantic markers to all lemmas identified as related to the domain. For this task, we used both our corpus (subsection 2.1) and the Serbian WordNet (subsection 2.3), as described in (Vujičić Stanković et al., 2014). When adding new entries we took care about language variants or pronunciation, e.g. Ekavian *belo vino* and Ijekavian *bijelo vino* ‘white wine’, so they were added into dictionaries no matter which form was actually occurring in the corpus. The set of markers is presented in Table 1. As a result, our e-dictionary now has 2,923 lemmas from the culinary domain – 1,607 simple lemmas and 1,316 compound lemmas.

2.3. Serbian WordNet

The development of WordNet for Serbian started in 2001 as a part of the BalkaNet Project.[6] As part of this project, EuroWordNet, corresponding to Princeton WordNet 2.1 (Fellbaum, 2010), was expanded by adding Balkan languages: Bulgarian, Greek, Romanian, Serbian, and Turkish. In 2004, at the end of the BalkaNet project, the Serbian WordNet (SWN) contained 7,000 synsets (Tufis et al., 2004). In the years that followed, the development continued, primarily on a voluntary basis. At present, the SWN is related to the Princeton WordNet 3.0 (PWN) and contains more than 21,200 synsets. The culinary domain is one of the domains that has been systematically filled – some characteristic branches in the hypernym/hyponym hierarchy were transferred from PWN to SWN by volunteering students and then used to automatically fill the gaps in Serbian e-dictionaries; and vice versa, lemmas from a Serbian e-dictionary and their culinary markers were used to fill the gaps in the SWN with Serbian-specific concepts (for more details see (Vujičić Stanković et al., 2014)). As a results of this procedure, the SWN has around 1,800 culinary concepts today, almost 550 of which are Serbian-specific concepts.[7]

2.4. Serbian Named Entity Recognition System

The Serbian Named Entity Recognition (NER) system is a handcrafted rule-based system that relies on comprehensive lexical resources for Serbian implemented in Unitex[8] (Krstev et al., 2013). It recognizes most major types of NEs: names of persons, locations and organizations, temporal expressions, and numeric expressions, including measures, money, amount, and percentage. For recognition of some types of named NEs, e.g. personal names and locations, e-dictionaries and the information in them are crucial; for others, like temporal expressions, local grammars in the form of FSTs that try to capture a variety of syntactic forms in which a NE can occur had to be developed. However, for all of them, local grammars were developed that use the wider context to disambiguate ambiguous occurrences, as much as possible. The latest version of the Serbian NER system is organized as a cascade of transducers, which means that several FSTs are applied on a text, one after the other. Each of them recognizes some sub-type of NEs, adds an

[4] Recepti: http://www.recepti.com/.
[5] Letters *č* and *ć* are used as *c*, *ž* as *z*, *š* as *s*, while *đ* is replaced by *dj*.
[6] BalkaNet: http://www.dblab.upatras.gr/balkanet/index.htm.
[7] SWN: http://resursi.mmiljana.com/Default.aspx.
[8] Unitex: http://www.igm.univ-mlv.fr/~unitex/.

appropriate tag to a text, which the FSTs applied subsequently can use. The use of cascades enables, among other things, the distinction between amount expressions and other expressions that use numerals, like measurement expressions.

Measurement and amount expressions, and to some extent temporal expressions are the most interesting for application to a corpus of culinary recipes. Our NER system recognizes the measurement expressions in which metric and U.S. units are used (in the form of simple words, compounds, and abbreviations) and a count of units of measure is expressed by numerals consisting of digits, words, and their combination. The recognized expressions represent either exact values, ranges of values or approximate values. One example is: *parče tvrdog sira od oko 100-150g* 'a piece of hard cheese about 100-150g'. Our NER system recognizes as amount expressions the phrases in which a numeral is followed by a count noun (possibly preceded by one or more adjectives) that agrees with it in the values of grammatical categories.

The evaluation results of our NER system against a news corpus were very good: precision 0.98, recall 0.94 and F-measure 0.96 for all NEs measured in tokens, precision 0.96, recall 0.88 and F-measure 0.92 measured in types. For measurement expressions precision was 0.99, recall 0.97 and F-measure 0.98 measured in tokens, while for types it was: precision 0.97, recall 0.94 and F-measure 0.96 (Krstev et al., 2013). However, there were not many such expressions in the analyzed corpus, only 289 of them in a 155,000 words from corpus. Our NER system recognized 48,531 measurement expressions and 65,749 amount expressions in our recipe corpus. We have not yet performed an evaluation on this new text type, but we expect that the performance is not as good.

3. Units of Measure in the Culinary Domain

One characteristic of all the recipes is extensive use of measurement expressions. Full understanding of these expressions is crucial for culinary professionals as "food costing, recipe size conversion, recipe development, and cost control" depend on it (Blocker & Hill, 2007). Moreover, it helps to calculate the quantity of food that should be prepared in order to obtain portions of the right size, because most foods shrink during preparation (Jones, 2008). Kitchens in different environments (restaurants, schools, hospitals, etc.) have special considerations regarding quantities and nutrition values (Edelstein, 2008). In their everyday life, people want to calculate the calories in the food they are preparing.

In order to achieve this, it is necessary to know what the units of measure are and how they relate to each other. The list of units of measure used in cooking given in (Edelstein, 2008) includes: units of length, volume and mass (metric and U.S. units, and their rates), temperature (Celsius and Farenheit), as well as the relation between standard scoop and can sizes. In (Jones, 2008), count as a unit of measure is listed as well. Blocker & Hill (2007) divide measure units into customary (such as graduated measures and nested measuring cups) and proper measures.

Culinary recipes written by users for other users (and not professionals for other professionals) are specific in the use of units of measure. Count and standard units of measure are used together with many informal units. As a preliminary step in our research, we have analyzed our corpus in order to obtain a general understanding of the units of measure used in the Serbian recipes. For that purpose, we used the tools described in subsection 2.4.

First we turned to standard units of measure. As expected, U.S. units of measure – *inč* 'inch', *unca* 'ounce', *stepen Farenhajta* 'degree Farenheit', etc. – are not used at all. As far as units of length are concerned, only centimeters are used, usually in the part of the recipe that describes the procedure: *Testo razviti na 1 cm debljine* 'Roll out the dough to become 1 cm thick', *Pleh veličine 20cm x 28cm podmazati* 'Oil the pan size 20cm x 28cm'. Centimeters are used only occasionally in the part of recipes that lists ingredients: *7 kotleta debljine oko 2 cm* '7 chops around 2 cm thick', *Jedan komad rebara širok 10 do 20 cm* 'One piece of ribs 10 to 20 cm wide'.

Degrees Celsius are the only measure of temperature, used, although the closer description *Celzijus* is rarely mentioned – only six times in our corpus: *Ugrejati pećnicu na 200 stepeni Celzijusa* 'Warm the oven to 200 degrees Celsius'. This unit of measure is predominantly used to describe the preparation phase, and only a few times to describe preservation of food: *Idealna je temperatura čuvanja oko 10 stepeni C* 'Ideal storage temperature is 10 degrees C'.

Finally, for describing food preparation units of time are used as well: *minut* 'minute', *sat*, *čas* 'hour', *dan* 'day': *Koru sušiti 100 minuta* 'Dry thin dough for 100 minutes', *Čuvajte ga u frižideru 2-3 dana* 'Keep it in refrigerator 2-3 days'.

As can be expected, units of counting are frequently used as a unit of measure, either to designate an exact quantity – *2 velika krompira* '2 big potatoes', *tri cijela jajeta* 'three whole eggs' – or as an approximate quantity – *nekoliko crnih maslinki* 'a few black olives'.

At this moment, we are interested only in the units of measure specifying the ingredients used in recipes. We observe that this information is often expressed by units of measure that are used more or less informally and are not listed in professional manuals; however, they have to be taken into consideration in order to accomplish the tasks previously mentioned (e.g. to automatize conversion from approximate measures to standard measures).

4. Approximate Units of Measure in the Culinary Domain

Our first goal was to produce an extensive list of the approximate units of measure that are used in the culinary domain. In order to do that, we have used all the resources for Serbian described in the previous section.

Our first task was to retrieve the approximate units of measure from our culinary corpus. To achieve this, we had to distinguish between count and uncount units of measure. In the latter case we have taken the following approach:

> *If a noun is preceded neither by a numeral nor by a unit of measure and is followed by a noun in the genitive case that refers to food (and possibly preceded by an adjective in the corresponding case, gender, number and animacy) then it can be said it refers to an uncount unit of measure.*

Only a few were found in the produced concordances: *prstohvat* and *na vrh noža*, both meaning a very small

amount ‘a pinch of’. For retrieving count approximate units of measure, we have taken the following approach:

> *If an amount expression is followed by a noun in the genitive case referring to some kind of food (and possibly preceded by an adjective in the corresponding case, gender, number and animacy) then we can presume that the noun used in the amount expression refers to a count approximate unit of measure.*

Our goal was to retrieve as many expressions as possible that contain approximate units of measure; however, we were not aiming at comprehensiveness that can result in too many false retrievals. Thus, we deliberately omitted the cases where amount expressions were not used at all. Our experiment with proper units of measure showed that units of measure in the culinary domain are seldom used without numerals – actually, just 16 such cases were found, e.g. *decilitar pavlake* ‘a deciliter_of cream’. We can safely presume that we will retrieve the majority of the approximate units of measure by following our approach. FST modelling it is shown in Figure 1.

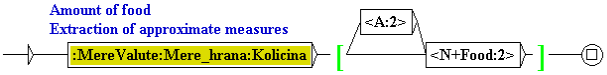


Figure 1. A FST that retrieves approximate units of measure. Agreement conditions are left out for simplicity purposes.

The FST in Figure 1 retrieved 15,521 lines of concordances; a few lines are shown in Figure 2. Only the candidates for units of measure can form part of concordance keywords, which facilitates the inspection of a large number of candidates and concordance lines. This is made possible by the use of contexts in graphs – the noun to which a unit of measure applies is used for retrieval but is not part of a keyword (green brackets in Figure 1). The same goes for numerals that are restricted to a context in the sub-graph **Kolicina** (the yellow box). The pattern <N+Food:2> retrieves all nouns in the genitive case related to food – both simple words (*vinobran* ‘potassium metabisulfite’) and compounds (*mladi luk* ‘fresh onion’).

3	**vezice**	crnog mladog luka
3	*small bunches*	*fresh onion*
jednu	**vezicu**	iseckanog peršunovog lista
one	*small bunch*	*chopped parsley leaves*
1	**vezu**	seckanog peršunovog lista
1	*bunch*	*chopped parsley leaves*
½	**vrećice**	praška za pecivo
½	*small pack*	*baking powder*
1	**vršna kašičica**	praška za pecivo
1	*peak small spoon*	*baking powder*
dva	**zrna**	suvog grožđa
two	*grains*	*raisins*
8-10	**zrnaca**	crnog bibera
8-10	*small corns*	*black papper*

Figure 2. A sample of the produced concordance lines.

The produced concordances were further analyzed by a volunteering student whose task was to select the candidates that represent approximate units of measure and mark those that are synonymous with other units of measure and/or are used only with some particular kind of food. As a result of this process, we obtained 106 approximate units of measure – 96 simple words and 10 compounds.

5. Approximate Measures Ontology and its Relation to Serbian Lexical Resources

A number of different ontologies of quantities and units of measure have been developed for different domains. For example, units of measurement ontology for biological and biomedical domains,[9] OASIS Quantities and Units of Measure Ontology Standard[10] for use across multiple industries, EngMath[11] for mathematical modeling in engineering, or Quantities, Units, Dimensions and Data Types Ontologies[12] and Ontology of Units of Measure (Rijgersberg et al., 2013) for a vast variety of quantitative research purposes, etc. The characteristic feature of these ontologies is that their scope is limited to formal measures, most frequently based on technical standards. Our goal is to develop an ontology for the informal measures specific to the culinary domain discussed in the above sections.

In order to enable semantic tagging of the approximate units of measure in the culinary domain, we modeled the OWL ontology. The ontology was modeled in the OWL 2 web ontology language[13] using the Protégé 4.3 tool,[14] because it makes it possible to establish a connection between classes and instances.

As to the discussed observations about the approximate measures in culinary recipes, we chose to use the introduced semantic categories as ontology classes, and the extracted units as instances. On the basis of the approximate units of measure extracted from our corpus, we introduced the following sub-classes of the top class *PribliznaMera* ‘ApproximateMeasure’: *Kontejner* ‘Container’, *Porcija* ‘Portion’, *DeoOd* ‘PartOf’, *Celina* ‘Whole’, and *Skupina* ‘Set’. Additionally, we proposed the object relationship property *jeManja* ‘isSmaller’ and the inverse property *jeVeca* ‘isBigger’ to signify that an approximate unit of measure is a smaller or bigger unit than another one from the same class. These classes with some instances from the class *Skupina* ‘Set’ are shown in Figure 3: *vezica* ‘small bunch’, *veza* ‘bunch’, *šaka* ‘handful’, *red* ‘row’; and the relationship property *jeManja*: *vezica jeManja veza* ‘small bunch *isSmaller* bunch’.

The analysis of concordances revealed that some approximate units of measure are used only for some particular kinds of food (or a restricted set), like *čen belog luka* ‘clove of garlic’ and *ploča lisnatog testa* ‘plate of puff pastry’ or *ploča lazanji* ‘plate of lasagna’, which is enforced in our ontology by introducing appropriate data properties *jeJedino* ‘isOnly’ and *jeIskljucivo*

[9] Units of measurement ontology: http://www.obofoundry.org/cgi-bin/detail.cgi?id=unit.
[10] OASIS QUOMOS: https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=quomos.
[11] EngMath: http://www-ksl.stanford.edu/knowledge-sharing/papers/engmath.html.
[12] QUDT: http://www.qudt.org/.
[13] OWL 2: http://www.w3.org/TR/owl2-overview/.
[14] Protégé: http://protege.stanford.edu/.

'isExclusively'. On the other hand, there are instances that belong to more than one class. Such is the case with *zrnce* that belongs to the class *PartOf* meaning 'grain' when referring to *biber* 'pepper' or *mak* 'poppy' and to the class *Portion* meaning 'small spherical amount of' when referring to *puter* 'butter'.

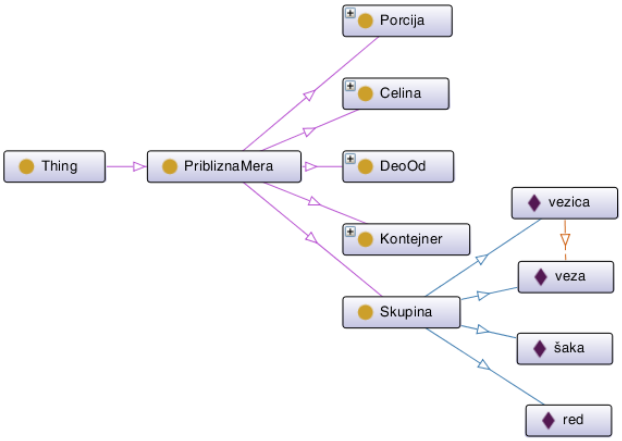


Figure 3. The hierarchy of approximate measures ontology classes, some instances, and their relationships.

Another very important aspect in ontology development is the possibility to designate that two or more instances refer to the same object. For example, *čen* and *češanj* '*clove of garlic*', and *štangla* and *rebro* 'bar of chocolate' should be treated as the same unit in culinary recipes. In the Serbian language, *šoljica za kafu*, *kafena šoljica*, *šoljica* and *kafena šolja* are different expressions for 'coffee cup'. It is sufficient to designate that information and the property of one of these instances, e.g. *šoljica jeManja šolja* 'cup *isSmaller* mug' in the ontology, for the reasoner to infer that the same is true for the other three instances (see Figure 4).

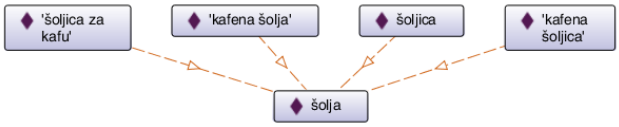


Figure 4. The same instances to which the property *isSmaller* is assigned.

Semantic marker	Description	Number of instances
+MesApp	approximate measure	106
+Cont	container (e.g. *supena kašika* 'soup spoon'	33
+Por	portion (e.g. *kriška* 'slice')	33
+Part	part of (e.g. *glavica* 'head')	30
+Wh	whole (e.g. *štapić* 'stick')	7
+Set	set (e.g. *veza* 'bunch')	4

Table 2. Semantic markers for approximate units of measure, typical representatives of classes and the number of instances in classes (some measures are in more than one class).

The ontology contains 7 classes, two object properties, two data properties, and 106 individuals. The ontology classes were used in the creation of a domain-specific e-dictionary of approximate units of measure. The new semantic markers and some representative instances from the classes are presented in Table 2.

Finally, we manually checked all the selected approximate units of measures against the SWN, and all those not already in it were added. This was not a straightforward task, because approximate measures in the culinary domain do not have a particular place in the PWN, and thus they do not have it in the SWN either. The only exception, to a certain extent, is 'containerful'. During this process, some units of measure were moved from one class to another that better corresponded to the PWN. For instance, *šaka* 'handful' was originally put in the class *Set*, but was afterwards moved to the class *Container* (because, 'handful' is a hyponym of 'containerful' in the PWN).

The work we have done is fully justified by the data presented in Table 3. In our culinary corpus, we have counted the expressions that use the units of measure applied to nouns representing some kinds of food by using the appropriate graphs. We cannot give an estimate of recall, but precision is very high (around 100% for the first column).[15] It can be seen that almost 45% of these expressions use approximate units of measure.

Units	With numerals	Without numerals	Total
Standard units	12,966	16	12,982
Approximate units	7,431	2,933	10,364

Table 3. Statistics of the use of units of measure in our culinary corpus.

Although this kind of knowledge could be, to some extent, represented in e-dictionaries and semantic networks, ontologies are much more suitable for useful reasoning. Moreover, the contribution of this ontology is the possibility of its integration in a comprehensive culinary domain recipe ontology on which we are currently working. To be more specific, in most cases, the culinary recipe structure is as follows: the name of the recipe (i.e. the meal that is in the focus of the recipe), the name of the author of the recipe, the part with listed ingredients that are required for recipe preparation together with the quantities, preparation description (usually listed in the steps that give a detailed account of the necessary utensils and way of preparation directions), and additional information like a summary of the recipe's nutritional values, preparation time or the level of preparation difficulty. Through detailed analysis of each of these parts, we came to a conclusion that it is necessary to develop a number of ontologies suitable for individual parts, which will later be integrated into a comprehensive culinary domain recipe ontology to represent the knowledge of the culinary domain.

[15] The produced concordances can be inspected at: http://poincare.matf.bg.ac.rs/~stasa//concordances/.

6. Conclusion

Our job is not yet finished, because there are still parts of the food ontology, e-dictionaries and WordNet that have to be filled. One major part still missing is related to the ways of preparation of food. However, the parts already developed can help in this. For instance, adjectives (and verbs) related to food preparation can be retrieved using the following procedure:

> *If an adjective derived from a verb past participle is preceded by numeric expressions with units of measure (standard and approximate) and followed by a noun in the genitive case that refers to food (and possibly preceded by an adjective in the corresponding case, gender, number and animacy) then it can be an adjective referring to a way of preparation of food.*

A graph developed following this approach retrieves with an almost 100% precision 1,805 concordance lines from our corpus related to adjectives (and the corresponding transitive verbs) we are looking for. From these, we have selected 85 adjectives and the corresponding verbs related to the culinary domain that vary from very general ones, like *pripremljen* 'prepared' and *pripremiti* 'prepare' to very specific ones like *pošećeren* 'sugared' and *pošećeriti* 'add sugar'. By these verbs as seeds for retrieval of more verbs and by modelling more procedures like this, we plan to prepare an exhaustive list of adjectives and verbs related to the culinary domain.

As was discussed in the previous section, we also plan to develop different ontologies like foodstuffs ontology, food product ontology, kitchen utensil ontology etc., in our future work in order to integrate all of them in one comprehensive culinary ontology and test in various applications.

7. Acknowledgements

We would like to thank the following PhD students at the Faculty of Philology, University of Belgrade for their help in enhancing the SWN with synsets from the culinary domain: Biljana Đorđević, Jelena Andonovska and Katarina Stanišić. This research was conducted as part of the project no. 178006, financed by the Serbian Ministry of Science.

8. References

Ahnert, S.E., 2013. Network analysis and data mining in food science: the emergence of computational gastronomy. In *Flavour* **2**:4. URI: http://dx.doi.org/10.1186/2044-7248-2-4.

Badra, F., R., Bendaoud, R., Bentebibel, P.A., Champin, J., Cojan, A., Cordier, S., Després et al., 2008. Taaable: Text Mining, Ontology Engineering, and Hierarchical Classification for Textual Case-Based Cooking. In *9th European Conference on Case-Based Reasoning-ECCBR 2008, Workshop Proceedings*, 219-228.

Batista, F., Pardal, J.P., Vaz Nuno Mamede, P., and R. Ribeiro, 2006. Ontology construction: cooking domain. *Artificial Intelligence. Methodology, Systems, and Applications* 4183 (2006): 213-221.

Blocker, L., and J. Hill, 2007. *Culinary Math*. John Wiley & Sons.

Cantais, J., Dominguez, D., Gigante, L., Laera, and V. Tamma, 2005. An example of food ontology for diabetes control. In *Proceedings of the International Semantic Web Conference 2005 workshop on Ontology Patterns for the Semantic Web*.

Chakkrit, S., and M. Bruckner, 2008. FOODS: a food-oriented ontology-driven system. In *Digital Ecosystems and Technologies. DEST 2008. 2nd IEEE International Conference on*. IEEE.

Courtois, B., L. M., Silberztein, et al., 1990. Dictionnaires électroniques du français. *Langue française*, 87(1):3–4. Armand Colin.

Dufour-Lussier, V., F., Le Ber, J., Lieber, T., Meilender, and E. Nauer, 2012. Semi-automatic annotation process for procedural texts: An application on cooking recipes. *arXiv preprint arXiv:1209.5663*.

Edelstein, S., 2008. *Managing Food and Nutrition Services: For the Culinary, Hospitality, and Nutrition Professions*. Jones & Bartlett Learning.

Fellbaum, C.. 2010. *WordNet*. Springer.

Jones, T., 2008. *Culinary Calculations: Simplified Math for Culinary Professionals*. John Wiley & Sons.

Kim, E., 2012. Korean Food Ontology. URL: http://kr-med.org/icbofois2012/fois/posters/kim_a4.pdf.

Krstev, C., 2008. *Processing of Serbian – Automata, Texts and Electronic Dictionaries*. Faculty of Philology, University of Belgrade.

Krstev, C., and D. Vitas, 2006. Finate State Transducers for Recognition and Generation of Compound Words. In *Proceedings of the 5th Slovenian and 1st International Conference Language Technologies*, IS-LTC 2006, Ljubljana, Slovenia, October, 2006, eds. T. Erjavec and J. Žganec Gros, 192-197, Institut "Jožef Stefan".

Krstev, C., I., Obradović, M., Utvić, and D. Vitas, 2013. A system for named entity recognition based on local grammars. *Journal of Logic Computation* 24(2): 473-489, Oxford Journals, doi:10.1093/logcom/exs079.

Mori, S., T., Sasada, Y., Yamakata, and K. Yoshino, 2012. A machine learning approach to recipe text processing. In *Computers workshop (CwC)*:29.

Nedovic, V., 2013. Learning recipe ingredient space using generative probabilistic models. In *Proceedings of International Joint Conference of Artificial Intelligence Workshops*, 13-18.

Pizzuti, T., G., Mirabelli, M.A., Sanz-Bobi, and F. Goméz-Gonzaléz, 2014. Food Track & Trace ontology for helping the food traceability control. *Journal of Food Engineering* 120 (2014): 17-30.

Rijgersberg, H., van Assem, M., and J. Top, 2013. Ontology of units of measure and related concepts. *Semantic Web 4*, no. 1 (2013): 3-13.

Tufis, D., D., Cristea, and S. Stamou, 2004. BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. *Romanian Journal of Information science and technology*, 7(1-2):9-43.

Vujičić Stanković, S., C., Krstev, and D., Vitas, 2014. Enriching Serbian WordNet and Electronic Dictionaries with Terms from the Culinary Domain. In *Proceedings of the seventh Global Wordnet Conference*.

Wiegand, M., B., Roth, and D. Klakow, 2012. Knowledge Acquisition with Natural Language Processing in the Food Domain: Potential and Challenges. In *Proceedings of the ECAI-Workshop on Cooking with Computers (CWC)*: 46-51.

Avtomatska razširitev in čiščenje sloWNeta

Darja Fišer,* Benoît Sagot†

* Oddelek za prevajalstvo, Filozofska fakulteta
Aškerčeva 2, 1000 Ljubljana, Slovenija
† UMR-I ALPAGE, UFR de Linguistique de l'Université Paris Diderot
Case 7003, 75205 Paris Cedex 13, France

Povzetek

V prispevku predstavljamo jezikovno neodvisno in avtomatsko razširitev wordneta z uporabo heterogenih že obstoječih jezikovnih virov, kot so strojno berljivi slovarji, vzporedni korpusi in Wikipedija. Pristop, ki ga preizkusimo na slovenščini, upošteva tako eno- kot večpomenske besede, splošno in specializirano besedišče, pa tudi eno- in večbesedne lekseme. Izluščenim besedam enega ali več pomenov pripišemo s pomočjo klasifikatorja, ki temelji na naboru različnih značilk, predvsem pa na distribucijski podobnosti. V naslednjem koraku s pomočjo distribucijskih informacij, izluščenih iz velikega korpusa, identificiramo in odstranimo zelo dvomljive kandidate. Avtomatska in ročna evalvacija rezultatov pokaže, da uporabljeni pristop daje zelo spodbudne rezultate.

Automatic extension and cleaning of sloWNet

In this paper we present a language-independent and automatic approach to extend a wordnet by recycling different types of already existing language resources, such as machine-readable dictionaries, parallel corpora and Wikipedia. The approach, applied to Slovene, takes into account monosemous and polysemous words, general and specialized vocabulary as well as simple and multi-word lexemes. The extracted words are assigned one or several synset ids based on a classifier that relies on several features including distributional similarity. In the next step we also identify and remove highly dubious (literal, synset) pairs, based on simple distributional information extracted from a large corpus in an unsupervised way. Automatic and manual evaluation show that the proposed approach yields very promising results.

1. Uvod

Avtomatski pristopi k izdelavi wordnetov so se uveljavili, ker je ročna izdelava časovno preveč potratna, da bi bila uresničljiva v večini raziskovalnih scenarijev, prav tako pa so avtomatske in polavtomatske metode že dale zadovoljive rezultate v številnih projektih, kot so npr. EuroWordNet, BalkaNet in Asian WordNet Vossen 1999, Tufis 2000, Sornlertlamvanich 2012). Pristop, predstavljen v tem prispevku, poleg avtomatske izdelave odlikuje predvsem dejstvo, da zanjo niso potrebni nobeni specializirani ali kompleksni algoritmi ali orodja, ki obstajajo samo za jezikovno-tehnološko najbolje podprte jezike, jezikovno odvisna pravila ali dragi leksikalni viri.

Z izkoriščanjem že obstoječih heterogenih jezikovnih virov, kot so strojno berljivi slovarji, vzporedni korpusi in Wikipedija, s predstavljenim pristopom maksimiziramo količino izluščenih leksikalnih informacij iz vsakega uporabljenega vira. Za razliko od večine sorodnih raziskav, ki temeljijo zgolj na Wikipediji, lahko s tem pristopom zajamemo vse besedne vrste, ne zgolj samostalnikov. Pristop je celovit tudi v smislu obravnave tako eno- kot večpomenskih, pa tudi eno- in večbesednih leksemov. Predstavljeni pristop je nadgradnja razvoja prvih različic slovenskega wordneta (Fišer in Sagot 2008), v okviru katere smo izboljšali tako obseg kot tudi natančnost razširjenega vira.

V 2. razdelku povzamemo sorodne raziskave, v 3. opišemo luščenje kandidatov za razširitev sloWNeta s pomočjo klasifikatorja, v 4. pa predstavimo filtriranje nezanesljivih kandidatov z uporabo načel distribucijske semantike. V 5. razdelku rezultate ročno in avtomatsko ovrednotimo, prispevek pa sklenemo z diskusijo in načrti za prihodnost.

2. Sorodne raziskave

Avtomatski pristopi izgradnje wordneta se med seboj razlikujejo predvsem glede na vrsto vira, ki ga za gradnjo uporabljajo. Najstarejši pristopi so temeljili na strojno berljivih dvojezičnih slovarjih, ki so služili za prevajanje sinsetov v Princeton WordNetu pod predpostavko, da slovarski prevodi predstavljajo isti pojem v ciljnem jeziku (Knight in Luk 1994, Yokoi 1995). Glavna ovira tega pristopa je, da dvojezični slovarji tipično niso pojmovno zasnovani, temveč slonijo na tradicionalnih leksikografskih načelih, kar otežuje razdvoumljanje slovarskih iztočnic. Pogosto je problematičen tudi njihov obseg oz. slovar za določen jezikovni par sploh ni na voljo.

Te slabosti so presegli različni pristopi, ki za prevajanje sinsetov uporabljajo dvo- in večjezične leksikone, izluščene iz vzporednih korpusov (Resnik in Yarowsky 1997, Fung 1995). Osrednja predpostavka tovrstnih pristopov je, da se pomeni večpomenskih besed v izvornem jeziku pogosto prevajajo v različne besede v ciljnem jeziku. Po drugi strani pa velja, da če se dve ali več besed v izvirniku prevajajo v isto besedo v ciljnem jeziku, imajo le-te pogosto skupne pomenske elemente. Posledično je mogoče identificirati pomenske razlike polisemnih besed oz. besede z enakim pomenom združiti v sopomenske nize, kar so dokazali Dyvik (2002), Ide idr. (2002) in Diab (2004).

Tretja skupina pristopov, ki so postali popularni v zadnjih nekaj letih, pa za iskanje prevodnih ustreznic uporablja Wikipedio, obsežno spletno enciklopedijo, ki v številnih jezikih nastaja s sodelovanjem zainteresiranih uporabnikov svetovnega spleta. Z njihovo pomočjo so raziskovalci zgradili nove wordnete s povezovanjem Wikipedijinih strani z najpogostejšim pomenom v Princeton WordNetu (Suchanek 2008), z izkoriščanjem

Wikipedijinih kategorij in drugih strukturnih informacij (Ponzetto in Navigli 2009) ter z luščenjem ključnih besed iz Wikipedijinih člankov (Reiter idr. 2008). Ruiz-Casado idr. (200) in Declerck idr. (2006) so s pomočjo modela vektorskega prostora mapirali Wikipedijine strani na WordNet. Najnaprednejši pristopi pa Wikipedijo in sorodne projekte, kot je npr. Wiktionary, uporabljajo za indukcijo wordnetov za številne jezike hkrati (de Melo 2009, Navigli in Ponzetto 2010, Navigli in Ponzetto 2012).

S pristopom, predstavljenim v pričujočem prispevku, smo za razširitev slovenskega wordneta uporabili vse vire, ki jih imamo na voljo: splošne in specializirane dvojezične slovarje, vzporedne korpuse in Wiki vire. Predstavljeni pristop je nadgradnja osnovne različice algoritma, za katere so bili uporabljeni isti viri (Erjavec in Fišer 2006, Fišer in Sagot 2008), v kateri smo osnovni pristop izboljšali tako, da deluje tudi za luščenje prevodnih ustreznic večpomenskih besed, kar omogoči poln izkoristek virov, ki so na voljo, pri čemer visoko stopnjo natančnosti zagotavlja ponderiranje kandidatov za sinsete glede na izbrane značilke.

3. Razširitev sloWNeta

Motivacija za razširitev sloWNeta izhaja iz dejstva, da smo pri izdelavi prve različice besede iz vzporednega korpusa razdvoumili s pomočjo ostalih jezikov v korpusu, medtem ko smo iz slovarjev in Wikipedije zaradi pomanjkanja ustreznih večjezičnih ali strukturnih informacij uporabili le enopomenske lekseme (Fišer in Sagot 2008), s čimer je precejšen delež dragocenih leksikosemantičnih informacij ostal neizkoriščen. Pristop, ki smo ga uporabili za razširitev sloWNeta tudi s temi informacijami, in tako bistveno izboljšali njegovo pokritost, opisujemo v tem razdelku, v naslednjem pa predstavimo varnostni mehanizem, s katerim smo kljub razširitvi v sloWNetu zagotovili visoko stopnjo natančnosti.

3.1. Verjetnostni klasifikator

Predstavljeni pristop temelji na verjetnostnem klasifikatorju, ki za odločanje o razvrščanju neke besede iz dvojezičnega slovarja oz. Wikipedije v obstoječi sloWNet uporablja niz značilk. Učno množico za klasifikator smo izdelali tako, da smo za vse izluščene pare (literal, sinset) – se pravi besedo v določenem pomenu –, ki že obstajajo v prejšnji različici sloWNeta, privzeli, da so pravilni, za vse ostale pa, da so nepravilni. Tako izdelana učna množica seveda ni popolna, saj po eni strani kot ustrezne pare (literal, sinset) obravnava napake, podedovane iz avtomatično generirane prve različice sloWNeta, po drugi pa tudi povsem ustrezne pare (literal, sinset) obravnava kot napačne, samo zato, ker se v prejšnji različici niso pojavili. Naš cilj v predstavljeni raziskavi je identificirati ravno te in z njimi razširiti wordnet.

Uporabili smo klasifikator največje entropije megam (Daume 2004) in s pomočjo značilk, opisanih v naslednjem razdelku, v razširjen wordnet vključili vse pare (literal, sinset), ki presegajo eksperimentalno določen prag verjetnosti 0,1.

3.2. Izbor značilk

Najpomembnejša značilka modelira semantično bližino med nekim literalom in potencialnimi sinseti. Naj jo ponazorimo na primeru angleško-slovenskega para *organ-organ*:

- Angleški leksem *organ* se v PWN 3.0. pojavi v 6 različnih sinsetih, zato smo za ta dvojezični par tudi generirali 6 različnih kandidatov (literal, sinset).
- Naša naloga je, da ugotovimo, kateri od teh 6 kandidatov so ustrezni, torej v katere sinsete v sloWNetu je potrebno dodati slovenski literal *organ*.
- Semantično bližino slovenskega literala z vsakim od 6 možnih sinsetov smo izračunali tako, da smo za vsak sinset izdelali vektor, v katerega smo vključili vse literale iz tega slovenskega sinseta in iz vseh sinsetov, ki so od osrednjega oddaljeni največ dve koleni.
- Tako na primer potencialen sinset {*organ*, *pipe organ*} iz PWN predstavlja naslednji slovenski vektor: {*glasbilo*, *Anton Bruckner*, *glasbenik*, *Johann Sebastian Bach*, *pisalni*, *klavirska*, *harmonika*,...}.
- Podoben vektor smo za potencialno slovensko prevodno ustreznico *organ* zgradili s pomočjo programskega pakega SementicVectors (Widdows in Ferraro 2008) iz korpusa FidaPLUS in ju nato primerjali v skladu z načeli distribucijske semantike.
- Semantična podobnost potencialne slovenske ustreznice *organ* s sinsetom {*organ*, *pipe organ*} znaša le 0.021, medtem ko primerjava s sinsetom {*organ, a fully differentiated structural and functional unit in an animal that is specialized for some particular function*}, znaša 0.668, kar je tudi jezikovno ustrezna rešitev.

Poleg mere semantične podobnosti smo uporabili tudi nekatere druge značilke, kot so število vseh angleških iztočnic, ki imajo pripisano isto slovensko ustreznico, najnižja stopnja večpomenskosti med vsemi angleškimi iztočnicami s pripisano isto slovensko ustreznic, število virov, iz katerih smo dvojezični par izluščili, in dolžina kandidata za prevodno ustreznico. Kot je razvidno iz tabele 1, se je v skladu s pričakovanji kot najbolj relevantna značilka izkazala semantična podobnost, saj ima največjo utež. H končnemu rezultatu pozitivno prispevata tudi indeks najnižje stopnje večpomenskosti za angleške literale in število različnih angleških literalov, ki imajo pripisano isto prevodno ustreznico. Po drugi strani pa na verjetnost ustreznosti kandidata za določen sinset negativno vpliva število pojavnic v slovenskem literalu.

Značilka	Utež
Semantična podobnost	6,24
Št. virov	0,55
Št. ang. literalov z isto ustreznico	0,33
Min. večpomenskost za ang. literal	2,69
Št. besed v slo. literalu	-1,87
Vir: Wikipedija	0,92
Vir: ang. Wiktionary	0,27
Vir: slo. Wiktionary	-0,07
Vir: SpeciesWiki	0,10
Vir: ang-slo slovar	0,15
Vir: slo-ang slovar	0,79

Tabela 1. Modeli za razvrščanje novih kandidatov (literal, sinset), naučeni na osnovnem wordnetu.

3.3. Rezultati klasifikacije

Naučen model smo uporabili za klasifikacijo 685.633 kandidatov, ki smo jih izluščili iz dvojezičnih slovarjev, vzporednega korpusa in Wikipedije. Mejni prag 0,1 je preseglo 68.070 kandidatov. Med njimi je 5.056 (7 %) takšnih, ki so že obstajali v prejšnji različici sloWNeta, novih pa je 63.010 (93 %) kandidatov. To pomeni, da je 25.102 sinsetov, ki so bili doslej prazni, dobilo vsaj en literal. Razširjena različica ima tako 141 % več neprazniih sinsetov kot pred razširitvijo. Število parov (literal, sinset) pa je še večje, saj smo jih z razširitvijo dobili 82.721, kar pomeni povečanje za 244 %.

4. Filtriranje sloWNeta

Kljub spodbudnim rezultatom postopek ni popoln, zato novi sinseti vsebujejo tudi precej šuma. Tega smo želeli odpraviti z jezikovno neodvisnim korpusnim pristopom za detekcijo in filtriranje potencialnih napak v avtomatsko generiranih sinsetih, s čimer bi dobili čistejši in uporabnejši semantični leksikon.

Čiščenje temelji na metodah distribucijske semantike za merjenje semantične podobnosti med besedami (Lin idr. 2003), vendar naš cilj ni prepoznavanje najbolj sorodnih besed glede na sobesedilo, v katerem se pojavljajo, temveč izhajamo iz (nepopolnega) seznama sinonimov, na katerem iščemo tiste, ki nanj ne sodijo. To pomeni, da je naloga podobna tistim na področju leksikalne substitucije (Mihalcea idr. 2010), pri čemer nas najbolj zanimajo kandidati, ki so na seznamu rangirani najnižje. Poleg tega je tudi naše razumevanje sinonimije strožje, saj je naš cilj očistiti vse sinsete v avtomatsko generiranem wordnetu, za katerega je znano, da ima pomene zelo nadrobno razdelane. Za to nalogo je ključno, da je naše dojemanje polisemije prevodno motivirano. To pomeni, da ne glede na število sinsetov, v katerih se beseda pojavi, so za naše potrebe pomenske razlike relevantne le, če se v ciljnem jeziku leksikalizirajo različno. Pri tem naj poudarimo še, da smo čiščenje wordneta zaenkrat sicer izvedli za vse besedne vrste, a zgolj za enobesedne literale (saj je stopnja večpomenskosti za večbesedne literale zelo nizka, medtem ko je procesiranje večbesednih literalov zahtevnejše, zato jih v nadaljevanju ne omenjamo, čeprav smo jih izluščili, predvsem iz Wikipedije, in vključili v razširjen sloWNet).

Naš cilj je identificirati in izločiti najočitnejše napake v sinsetih, ki so se v njem pojavile zaradi napačne besedne poravnave vzporednega korpusa ali napačnega razdvoumljanja homonimov, saj ravno tovrstne napake najbolj znižujejo uporabno vrednost wordneta. Zato predstavljeni pristop temelji na preprosti tezi: leksemi (oz. pari (literal,sinset)) se v korpusi sopojavljajo s semantično povezanimi leksemi, ki so eksplicitno kodificirani s semantičnimi relacijami v wordnetu. Pristop je sestavljen iz dveh korakov:

- primerjava kontekstualne podobnosti leksemov v referenčnem korpusu s sinseti v wordnetu,
- globalna razvrstitev vseh parov (literal, sinset) glede na dobljene rezultate.

Za vsak par (literal, sinset) smo najprej zgradili vektor, v katerega smo vključili vse literale iz vseh sinsetov, ki so s ciljnim povezani z ročno izbranimi semantičnimi relacijami (hipernimija, hiponimija, holonimija, meronimija, derivacija) v razdaji 0-2. Nato smo za vsakega od teh literalov zgradili še kontekstni vektor iz referenčnega korpusa, v katerega smo vključili vse polnopomenske besede, ki se pojavijo v istem odstavku. V zgrajenih vektorjih smo nato primerjali stopnjo prekrivanja kontekstov z algoritmom, zelo podobnim Leskovemu, ki je klasična mera za razdvoumljanje večpomenskih besed s pomočjo slovarjev (Lesk 1986). Pare (literal, sinset) smo nato rangirali tako, da smo za vsak iteral, ki je povezan s sinsetom in se pojavi v istem odstavku, rezultat povečali za število pojavitev v korpusu, deljeno s številom sinsetov, v katerih se v wordnetu pojavi. S tem smo dali manjšo težo zelo polisemnim literalom. Rezultat smo nato normalizirali še s številom polnopomenskih besed v odstavku.

Če pristop ponazorimo na primeru literala *ikona*, opazimo, da se v sloWNetu pojavi v 4 sinsetih, med drugim tudi v teh dveh:

- *eng-30-07269916-n {icon}*; ikona je v tem primeru ustrezen prevod, v vektorju sinseta pa so naslednji povezanimi literali: *znak, točka, simbol, računalništvo...*,
- *eng-30-03931044-n {icon, ikon, image, picture}*; ikona v tem primeru ni ustrezen prevod, v vektorju sinseta pa so naslednji povezani literali: *fotografija, podoba, predstaviti, prikaz...*

V korpusu FidaPLUS se samostalnik *ikona* pojavi 3.488-krat. Seštevek vseh rezultatov za pojavitev besede ikona v korpusu za pravilen par *(ikona, eng-30-07269916-n)* glede na zgoraj omenjen vektor znaša le 1,02, medtem ko globalni rezultat za nepravilen par *(ikona, eng-30-03931044-n)* znaša 5,99, kar pomeni, da tak globalni rezultat ni učinkovit indikator napak v razširjenem wordnetu. Zato smo ga nadgradili tako, da smo ga normalizirali z vsoto vseh globalnih rezultatov za par (literal, sinset) za celotni sinset, kar meri prispevek določenega literala med vsemi literali v sinsetu. Prispevek literala smo nato normalizirali še s številom pojavitev tega literala v korpusu, dobljen rezultat pa je hkrati tudi končni rezultat.

V našem prejšnjem primeru globalni rezultat za celotni sinset za pojem *eng- 30-07269916-n* znaša 1,02, za pojem *eng-30-03931044-n* pa 234, medtem ko prispevka literala za ta sinseta znašata 1 in 0,026. Normaliziran prispevek literala oz. končni rezultat za par *(ikona, eng-30-07269916-n)* znaša 0,287 za par *(ikona, eng-30-03931044-n)* pa le 0,007, s čimer je napačen kandidat v razširjenem wordnetu ustrezno identificiran.

Glede na izmerjeno 18-odstotno stopnjo napake razširjenega sloWNeta, smo za mejni prag pri filtriranju sloWNeta določili takšno vrednost, ki iz njega izloči primerljiv delež parov (literal, sinset). Ta znaša $4 \cdot 10{-}6$ in iz sloWNeta izloči 12,578 parov oz. tretjino vseh identificiranih potencialnih napak.

5. Vrednotenje rezultatov

Vrednotenje rezultatov smo opravili ročno in avtomatsko, pri čemer smo ročno evalvirali razširjen sloWNet in detekcijo napak, za avtomatsko vrednotenje pa smo izdelan vir primerjali z manjšim zlatim standardom ter z avtomatsko generiranima večjezičnima viroma Universal WordNet (de Melo 2009) in BabelNet 2.0 (Navigli in Ponzetto 2010, Navigli in Ponzetto 2012), ki vsebujeta tudi slovenščino.

5.1. Ročno vrednotenje razširjenega sloWNeta in avtomatske detekcije napak

Ročno vrednotenje razširjenega sloWNeta smo opravili na vzorcu 100 naključnih parov (literal, sinset) iz razširjenega sloWNeta za vsako besedno vrsto. Na podlagi teh rezultatov lahko podamo skupno oceno celotnega wordneta, in sicer tako, da rezultate za posamezno besedno vrsto ponderiramo z relativnim številom parov (literal, sinset). Kot je razvidno iz tabele 2, smo najvišjo stopnjo natančnosti izmerili za prislove (96%), najnižjo za glagole (59%), medtem ko skupni rezultat za razširjen sloWNet 3.0 znaša 82%.

Bes. vrsta	Σ parov	% parov	% pravilnih	% nepravilnih
sam.	55.383	67 %	87 %	13 %
prid.	12.438	15 %	85 %	15 %
gl.	14.053	17 %	59 %	41 %
prislov	847	1 %	96 %	4 %
Skupaj	82.721	100 %	82 %	18 %

Tabela 2. Rezultati ročnega vrednotenja razširjenega sloWNeta.

Ročno vrednotenje detekcije napak v razširjenem sloWNetu smo opravili na 100 naključnih parih (literal, sinset), ki jih je algoritem prepoznal kot napačne. Stopnja natančnosti tega avtomatskega postopka je 64 %.

Glede na to, da smo za razširitev sloWNeta uporabili prag, s katerim smo iskali optimalno razmerje med priklicem in natančnostjo, smo v leksikon vnesli tudi nekaj šuma. Zato je spodbudno, da z detekcijo potencialnih napak identificiramo kandidate, med katerimi je 64 % dejanskih napak. Ne samo, da z algoritmom najdemo več napak, kot bi jih s pregledovanjem naključnih kandidatov, temveč algoritem uspešno izkoristi razširjeno mrežo, ki doslej ni bila na voljo.

5.2. Avtomatsko vrednotenje razširjenega sloWNeta

Avtomatska primerjava z ročno izdelanim zlatim standardom, izdelanim z ročno validacijo prve različice sloWNeta, ki je temeljila na srbskem wordnetu (Erjavec in Fišer 2006) in vsebuje osnovni nabor sinsetov, pokaže 70 % natančnost, kar je sicer manj, kot je pokazala ročna evalvacija v prejšnjem razdelku, a je treba poudariti, da zlati standard vsebuje predvsem osnovni nabor sinsetov, ki vsebujejo zelo splošno besedišče, za katerega je značilna visoka stopnja večpomenskosti, tako da je za ta segment besedišča avtomatska naloga bistveno težja.

Nekoliko drugačna evalvacija, ki ovrednoti predvsem pristop, s katerim smo sloWNet izdelali, pa je primerjava s sorodnimi leksikosemantičnimi viri, in sicer z BabelNetom (Navigli in Ponzetto 2010, Navigli in Ponzetto 2012) in Universal WordNetom (de Melo 2009). Čeprav vsi trije viri temeljijo na primerljivih virih, je bil za razliko od nas osnovni cilj UWN in BabelNeta izdelati večjezično semantično mrežo.

Zato je razumljivo, da je slovenski del UWN z 9.924 pari (literal, sinset) precej manjši od sloWNeta, ki jih vsebuje 82.721. Kot je razvidno iz tabele 3, 5.590 (56 %) od 9.924 slovenskih parov v UWN vsebuje tudi sloWNet 3.0.

Podrobni rezultati					
		so v BabelNetu 2.0		niso v BabelNetu 2.0	
		so v UWN	niso v UWN	so v UWN	niso v UWN
so v sloWNetu 3.0	št. parov	2.239	12.468	3.351	64.663
	natančnost	98 %	98 %	92 %	86 %
niso v sloWNetu 3.0	št. parov	901	116.367	3.433	-
	natančnost	100 %	70 %	72 %	-
Primerjava z BabelNetom 2.0					
		so v Babelnetu 2.0		niso v Babelnetu 2.0	
so v sloWNetu 3.0	št. parov	14.707		68.014	
	natančnost	98 %		86 %	
niso v sloWNetu 3.0	št. parov	117.257		-	
	natančnost	70 %		-	
Primerjava z UWN					
		so v UWN		niso v UWN	
so v sloWNetu 3.0	št. parov	5.590		77.131	
	natančnost	94 %		88 %	
niso v sloWNetu 3.0	št. parov	4.334		-	
	natančnost	78 %		-	
Pregled posameznih virov					
	sloWNet 3.0	BabelNet 2.0		UWN	
št. parov	82.721	131.964		9.924	
natančnost	88 %	73 %		87 %	

Tabela 3. Primerjava razširjenega sloWNeta z BabelNetom in UWN.

Po drugi strani pa sloWNet vsebuje še 77.131 (93 %) parov, ki jih v UWN ni. Za razliko od UWN BabelNet 2.0 vsebuje kar 131.964 parov. Med njimi jih je 14.707 (11 %) tudi v sloWNetu 3.0. 69.014 (82 %) parov iz sloWNeta ni v BabelNetu. Če primerjamo vse tri vire, kar 64.663 parov najdemo samo v sloWNetu, po drugi strani pa je samo 901 parov tako v BabelNetu in UWN, v sloWNetu pa manjkajo.

Za boljšo predstavo o kvaliteti primerjanih virov smo ročno ovrednotili po 50 naključnih parov (literal, sinset) za vsakega od zgoraj naštetih scenarijev. Skupna natančnost za sloWNet znaša 88 %, kar je primerljivo z UWN, venar je UWN precej manjši. Natančnost 64.663 parov, ki jih najdemo samo v sloWNetu, je 86 %, kar je bistveno več kot natančnost parov, ki so samo v UWN (72 %), in tistih, ki so samo v BabelNetu (72 %).

Pristopi, ki so bili uporabljeni za gradnjo teh treh virov, so komplementarni, saj so praktično vsi pari, ki jih vsebujejo vsi trije viri, pravilni (2,239), zelo kvalitetni pa so tudi pari, ki si jih delita po dva vira (92 %).

BabelNet, sicer zelo obširen vir, je manj zanesljiv, saj je pravilnih samo 70 % parov, ki jih ni v sloWNetu, v primerjavi z 78 % pari v UWN, ki jih ni v sloWNetu.

Napake v sloWNetu poleg napačne disambiguacije so povezane z zastarelim dvojezičnim slovarjem, ki vsebuje precej arhaičnih izrazov, napake v parih, ki jih najdemo samo v UWN in BabelNetu, pa večinoma tičijo v napačni normalizaciji, kot so neprevedeni ang. izrazi, naslovi strani v Wikipediji, ki niso literali (*Seznam Arheoloških Dob*), semantično ustrezne slovenske besede, a napačne besedne vrsto, v ženski obliki, se začnejo s številko, končajo s piko ali disambiguatorjem (npr. *Mars (bog)*). Glede na opravljeno primerjalno analizo ugotavljamo, da je sloWNet 3.0 najkvalitetnejši leksikosemantični vir za slovenščino, ki je trenutno na voljo.

6. Zaključek

V prispevku smo predstavili avtomatsko širitev in čiščenje slovenskega semantičnega leksikona sloWNet s pomočjo različnih že obstoječih več- in dvojezičnih virov, kot so dvojezični slovarji, vzporedni korpusi in Wiki viri. S širitvijo smo dragocene leksikosemantične informacije v njih izkoristili v največji možni meri, ne samo večpomenskih iz korpusa in enopomenskih iz slovarja in Wikipedije, kot je to bilo storjeno v gradnji prejšnje verzije sloWNeta. Za pripisovanje pravega pomena večpomenskim besedam smo si pomagali s klasifikatorjem, ki je za odločanje uporabljal različne značilke, predvsem pa distribucijsko podobnost. S širitvijo smo število nepraznih sinsetov v sloWNetu povečali s 17.817 na 42.919, število parov (literal, sinset) pa je z 24.081 poskočilo na 82.721 (+244 %).

Ročno in avtomatsko vrednotenje tako razširjenega wordneta pokaže 85 % natančnost in ima bistveno večji priklic v primerjavi s prejšnjo različico. Razširjen sloWNet smo naložili v prosto dostopno spletno orodje za brskanje, editiranje in vizualizacijo wordneta, sloWTool (Fišer in Novak 2011), s čimer je na voljo študentom, prevajalcem in drugim jezikoslovcem, celotna podatkovna zbirka pa je dostopna pod licenco Creative Commons BY-SA (s priznanjem avtorstva in deljenjem pod enakimi pogoji): http://nl.ijs.si/sloWNet.

V prihodnje želimo izboljšati luščenje leksikosemantičnih informacij iz Wikipedije in Wiktionarija, s čimer bi lahko v sloWNet dodali definicije in primere rabe. Pristop želimo razširiti tudi na primerljive korpuse (Fišer idr. 2012), ki jih je veliko lažje pridobiti s spleta kot vzporedne. Prav tako pa si pristop, ki se je že izkazal kot učinkovit za gradnji francoskega wordneta (Sagot in Fišer 2008), želimo preizkusiti še na hrvaščini.

Literatura

H. Daume III. 2004. *Notes on CG and LM-BFGS optimization of logistic regression*.

T. Declerck, A.G. Pérez, O. Vela, Z. Gantner, D. Manzano-Macho. 2006. Multilingual lexical semantic resources for ontology translation. *Zbornik konference International Conference on Language Resources and Evaluation (LREC '06)*. Genova, Italija.

M. Diab. 2004. The feasibility of bootstrapping an Arabic wordnet leveraging parallel corpora and an English wordnet. *Zbornik konference Arabic Language Technologies and Resources.*

H. Dyvik. 2002. Translations as semantic mirrors: from parallel corpus to wordnet. *Zbornik konference ICAME '02*. Gothenburg, Švedska.

T. Erjavec, D. Fišer. 2006. Building Slovene wordnet. *Zbornik konference International Conference on Language Resources and Evaluation (LREC '06).* Genova, Italija.

D. Fišer, B. Sagot. 2008. Combining Multiple Resources to Build Reliable Wordnets. *Zbornik konference Text, Speech and Dialogue (TSD '08)*. Brno, Češka.

D. Fišer, N. Ljubešić, O. Kubelka. 2012. Addressing polysemy in bilingual lexicon extraction from comparable corpora. *Zbornik konference International Conference on Language Resources and Evaluation (LREC '12)*. Istanbul, Turčija.

Fišer, D., Novak, J. 2011. Visualizing sloWNet. *Zbornik konference Electronic lexicography in the 21st century: new applications for new users (eLEX '11)*. Bled, Slovenija.

P. Fung. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. *Zbornik konference Association for Computational Linguistics*. Stroudsburg, PA, ZDA.

N. Ide, T. Erjavec, D. Tufis. 2002. Sense discrimination with parallel corpora. *Zbornik konference Association for Computational Linguistics, Workshop on word sense disambiguation: recent successes and future directions.* Stroudsburg, PA, ZDA.

K. Knight, S. K. Luk. 1994. Building a large-scale knowledge base for machine translation. *Zbornik konference Artificial intelligence*. Menlo Park, CA, ZDA.

D. Lin. 1998. An information-theoretic definition of similarity. *Zbornik konference Machine Learning*. Madison, Wisconsin, ZDA.

M. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *Zbornik conference 5th annual international conference on systems documentation (SIGDOC '86)*, 24-26.

G. de Melo, G. Weikum. 2009. Towards a universal wordnet by learning from combined evidence. *Zbornik konference Information and knowledge management, (CIKM '09)*.

R. Mihalcea, R. Sinha, D. McCarthy. 2010. Semeval-2010 task 2: Cross-lingual lexical substitution. *Zbornik delavnice 5th international workshop on semantic evaluation.*

R. Navigli, S. P. Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. *Zbornik konference Association for Computational Linguistics*, Uppsala, Švedska.

R. Navigli, S. P. Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193, 217–250.

R. Navigli, S. P. Ponzetto. 2009. Large-scale taxonomy mapping for restructuring and integrating wikipedia. *Zbornik konference Artifical intelligence (IJCAI '09)*. San Francisco, CA, USA.

N. Reiter, M. Hartung, A. Frank. 2008. A Resource-Poor Approach for Linking Ontology Classes to Wikipedia Articles. *Zbornik konference Semantics in Text Processing (STEP '08)*.

P. Resnik, D. Yarowsky. 1997. A perspective on word sense disambiguation methods and their evaluation. *Zbornik delavnice Tagging Text with Lexical Semantics: Why, What, and How?*. Washington, D.C., ZDA.

M. Ruiz-Casado, E. Alfonseca, P. Castells. 2005. Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. *Zbornik konference Advances in Web Intelligence*.

B. Sagot, D. Fišer. 2008. Building a free French wordnet from multilingual resources. *Zbornik konference Ontolex 2008*. Marakeš, Maroko.

V. Sornlertlamvanich. 2010. Asian wordnet: Development and service in collaborative approach. *Zbornik konference Global WordNet Association (GWC '10)*. Mumbaj, Indija.

F. M. Suchanek, G. Kasneci, G. Weikum. 2008. Yago: A large ontology from Wikipedia and WordNet. *Journal of Web Semantics 6(3)*, 203–217.

D. Tufis. BalkaNet – Design and Development of a Multilingual Balkan WordNet. 2000. *Romanian Journal of Information Science and Technology Special Issue* 7(1–2).

P. Vossen. 1999. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht.

D. Widdows, K. Ferraro. 2008. Semantic vectors: a scalable open source package and online technology management application. *Zbornik konference International Conference on Language Resources and Evaluation (LREC '08)*. Marrakech, Morocco.

T. Yokoi. 1995. The EDR electronic dictionary. *Zbornik konference ACM 38 (11)*, 42–44.

The slWaC 2.0 Corpus of the Slovene Web

Tomaž Erjavec[†] and Nikola Ljubešić[*]

[†]Dept. of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39 1000 Ljubljana
tomaz.erjavec@ijs.si
[*]Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
nikola.ljubesic@ffzg.hr

Abstract

Web corpora have become an attractive source of linguistic content, as they can be made automatically, contain varied text types of contemporary language, and are quite large. This paper introduces version 2 of slWaC, a web corpus of Slovene containing 1.2 billion tokens. The corpus extends the first version of slWaC with new materials and updates the corpus compilation pipeline. The paper describes the process of corpus compilation with a focus on near-duplicate removal, presents the linguistic annotation, format and accessibility of the corpus via web concordancers, and then investigates the content of the corpus using frequency profiling, by comparing its lemma and part-of-speech annotations with the first version of slWaC and with KRES, the reference balanced corpus of Slovene.

Korpus slovenskega spleta slWaC 2.0

Korpusi besedil zajetih s spleta so postali popularen vir jezikovnih vsebin, saj jih lahko zgradimo avtomatsko, vsebujejo pester nabor sodobnih besedilnih zvrsti in so zelo veliki. Prispevek predstavi drugo različico korpusa slWaC, spletnega korpusa slovenščine, ki vsebuje 1,2 milijarde pojavnic. Korpus dopolnjuje prvo različico slWaC z novimi besedili, pridobljenimi z izboljšanimi orodji za zajem. V prispevku opišemo proces izdelave korpusa s poudarkom na odstranjevanju podobnih vsebin, predstavimo jezikoslovno označevanje, format korpusa in njegovo dostopnost preko konkordančnika. Nato raziščemo vsebino korpusa s pomočjo frekvenčnega profila, kjer leme in oblikoskladenjske oznake druge različice korpusa slWaC primerjamo s prvo ter z referenčnim in uravnoteženim korpusom slovenščine KRES.

1. Introduction

With the advent of the web, a vast new source of linguistic information has emerged. The exploitation of this resource has especially gained momentum with the WaCky initiative (Baroni et al., 2009), which has popularised the concept of "Web as Corpus". It has also made available tools for compiling such corpora and produced large WaC corpora for a number of major European languages. Now such corpora are also being built for the so called smaller languages, such as Norwegian (Guevara, 2010), Czech (Spoustová et al., 2010) and Serbian (Ljubešić, 2014), moving the concept of a "large corpus" for smaller languages up to the 1 billion token frontier. As Web corpus acquisition is much less controlled than that for traditional corpora, the necessity of analysing their content gains in significance. The linguistic quality of the content is mostly explored through word lists and collocates (Baroni et al., 2009) while the content itself is explored using unsupervised methods, such as clustering and topic modelling (Sharoff, 2010).

For Slovene, a web corpus has already been built (Ljubešić and Erjavec, 2011). However, the first version of slWaC (hereafter slWaC_1) was rather small, as it contained only 380 million words. Furthermore, it contained domains from the Slovene top-level domain (TLD) only, i.e. only URLs ending with ".si" were harvested. In the meantime, hrWaC, the Croatian web corpus had already moved to version 2, touching the 2 billion token mark, and web corpora for Serbian and Bosnian were built as well (Ljubešić, 2014), all of them passing the size of slWaC_1, making it high time to move forward also with slWaC.

This paper presents version 2 of slWaC (hereafter slWaC_2) which tries to overcome the limitations of slWaC_1: it extends it with a new crawl, which also includes well known Slovene web domains from other TLDs, and introduces a new pipeline for corpus collection and cleaning, resulting in a corpus of 1.2 billion tokens with removed near-duplicate documents and flagged near-duplicate paragraphs.

The rest of the paper is structured as follows: Section 2 presents the corpus construction pipeline, Section 3 introduces the linguistic annotation of the corpus, its format and its availability for on-line concordancing, Section 4 investigates the content of the corpus, by comparing it to slWaC_1 and to the KRES balanced corpus of Slovene, while Section 5 gives some conclusions and directions for future work.

2. Corpus construction

2.1. Crawling

For performing the new crawl we used the SpiderLing crawler[1] with its associated tools for guessing the character encoding of a web page, its content extraction (boilerplate removal), language identification and near-duplicate removal (Suchomel and Pomikálek, 2012). The SpiderLing crawler has two predefined size ratio thresholds that control when a low-yield-rate web domain (concerning new text) is to be abandoned; we used the lower one which is recommended for smaller languages. As seed URLs we used the home pages of web domains obtained during the construction of slWaC_1 and additionally 30 well known Slovene web domains, which are outside the .si TLD.

[1] http://nlp.fi.muni.cz/trac/spiderling

The crawl was run for 21 days, with 8 cores used for document processing, which includes guessing the text encoding, text extraction, language identification and physical duplicate removal, i.e. removing copies of identical pages which appear under different URLs. After the first 14 days there was a significant decrease in computational load, showing that most of the domains had been already harvested and that the process of exhaustively collecting textual data from the extended Slovene TLD was almost finished.

After completing the crawling process, which already included document preprocessing, we merged the new crawl with slWaC$_1$. We added the old dataset to the end of the new one, thereby giving priority to new data in the following process of near-duplicate removal. It should be noted that the corpus can, in cases when the content has changed, contain two texts with the same URL but with different crawl dates.

2.2. Near duplicate removal

We performed near-duplicate identification both on the document and the paragraph level using the onion tool[2] with its default settings, i.e. by calculating 5-gram overlap and using the 0.5 duplicate content threshold. We removed the document-level near-duplicates entirely from the corpus, while keeping paragraph-level near-duplicates, labelling them with a binary attribute on the <p> element. This means that the corpus still contains the (near)duplicate paragraphs, which is advantageous for showing contiguous text from web pages, but if, say, language modelling for statistical machine translation were to be performed (Ljubešić and Toral, 2014), near-duplicate paragraphs can easily be removed.

The resulting size of the corpus (in millions of tokens) after each of the three duplicate removal stages is given in Table 1. We compare those numbers to the ones obtained on the Croatian, Bosnian and Serbian domains (Ljubešić, 2014), showing that the second versions of the corpora (hrWaC and slWaC), which merge two crawls obtained with different tools and were collected three years apart, show a smaller level of reduction (around 30%) at each step of near-duplicate removal, while the first versions of corpora (bsWaC and srWaC), obtained with SpiderLing only and in one crawl, suffer more data loss in this process (around 35-40%).

	PHYS	DOCN	PARN	R1	R2
slWaC 2	1,806	**1,258**	**895**	0.31	0.29
hrWaC 2	2,686	1,910	1,340	0.29	0.30
bsWaC 1	722	429	288	0.41	0.33
srWaC 1	1,554	894	557	0.42	0.37

Table 1: Sizes of the web corpora in millions of tokens after removing physical duplicates (PHY), document near-duplicates (DOCN) and paragraph near-duplicates (PARN), with the reduction ratio (R1 and R2) after the DOCN and subsequent PARN steps.

[2]https://code.google.com/p/onion/

2.3. Linguistic annotation

slWaC$_2$ was tagged and lemmatised with ToTaLe (Erjavec et al., 2005) trained on JOS corpus data (Erjavec and Krek, 2008). However, it should be noted that ToTaLe had been slightly updated, so in particular the tokenisation of slWaC$_1$ and slWaC$_2$ at times differs. The morphosyntactic descriptions (MSDs) that the words of the corpus are annotated with follow the JOS MSD specifications, however, these do not define a tag for punctuation. As practical experience has shown this to be a problem, we have introduced a punctuation category and MSD, named "Z" in English and "U" in Slovene.

3. Overview of the corpus

3.1. Size of the corpus

Table 2 gives the size of slWaC$_2$, for the included slWaC$_1$from 2011 and the new additions in 2014, and together. For each of the counted elements we also give the size of the complete corpus, i.e. after removing document near duplicates (DOCN from Table 1), and for the corpus which has also paragraph near duplicates removed (PARN).

slWaC$_2$	2011	2014	All
Domains	25,536	22,062	37,759
URLs	1,528,352	1,295,349	2,795,386
Pars	7,535,453	18,303,123	25,838,576
(PARN)	6,325,075	10,329,692	16,654,767
Sents	22,615,610	50,693,747	73,309,357
(PARN)	19,001,653	31,560,289	50,561,942
Words	360,273,022	718,332,186	1,078,605,208
(PARN)	301,547,669	465,780,456	767,328,125
Tokens	421,178,853	837,727,874	1,258,906,727
(PARN)	352,474,874	542,912,192	895,387,066

Table 2: Size of the slWaC 2.0 corpus.

Starting with the number of domains, it can be seen that the new crawl produced less domains than the first one, due to a large number (of the complete space of URLs) of static domains being removed in the physical deduplication stage (PHY). Nevertheless, the complete corpus has, in comparison to slWaC$_1$, about 12,000 new domains. Observing the URLs, we note that the new crawl gave somewhat less URLs than the old one, and that there is little overlap between the two, i.e. about 1%: 28,315 URLs are the same from both crawls, which means that their content has changed in the last three years (and are then in the corpus distinguished by having a different crawl date).

Starting the the number of paragraphs we give both the numbers for DOCN and PARN, with the reduction having been already expressed in Table 1, i.e. 29%. For paragraphs, sentences, words and tokens, the complete corpus is simply the sum of the items for each of the two crawls. The most important numbers are the sizes of the complete corpus in tokens, i.e. 1.25 billion words for the DOCN and 900 million for PARN, which makes the corpus almost as large as the largest corpus of Slovene to date, i.e. Gigafida.

3.2. Corpus format

The annotated corpus is stored in the so called vertical format, used by many concordancing engines. This is an XML-like format in that it has opening and closing or empty (structural) XML tags, but the tokens themselves are written one per line, with the first (tab separated) column giving the token (word or punctuation) itself, the second (in our case) its lemma (or, for punctuation, again the token), the third its MSD in English and the fourth the MSD in Slovene, as illustrated by Figure 1.

```
<text domain="www.cupradan.si"
      url="http://www.cupradan.si/"
      crawled="2014">
<gap extent="1000+"/>
<p type="text" duplicate="0">
<s>
*        *       Z         U
Izmed    izmed   Sg        Dr
vseh     ves     Pg-mpg    Zc-mmr
<g/>
,        ,       Z         U
ki       ki      Cs        Vd
boste    biti    Va-f2p-n  Gp-pdm-n
delili   deliti  Vmpp-pm   Ggnd-mm
video    video   Ncmsan    Sometn
...
```

Figure 1: Vertical format of the annotated slWaC$_2$.

The example also shows a few other features of the encoding. Each text is given its URL, the domain of this URL and the year (2011 or 2014) on which it was crawled. Boilerplate removal often deletes linguistically uninteresting texts from the start (and end) of the document, which is marked by the empty gap element, which also gives the approximate extent of the text removed. The paragraphs are marked by their type, which can be "heading" or "text", while the "duplicate" attribute tells whether the paragraph is a (near) duplicate of some other paragraph in the corpus, in which case its value is "1", and "0" otherwise. Finally, we also have the empty "glue" element g, which can be used to remove the space between two adjacent tokens in displaying the corpus.

3.3. Availability

The corpus is mounted under the noSketchEngine concordancer (Rychlỳ, 2007) installed at nl.ijs.si/noske. The concordancer allows for complex searches in the corpus, from concordances taking into account various filters, to frequency lexica over regular expressions.

We also make the corpus available for download, but not directly, mainly due to question of personal data protection. Namely, the corpus contains most of the Slovene Web, at least in the .si domain, so it also contains a lot of personal names with accompanying text. This is not such a problem with the concordancer, as simliar results on Web-accessible personal names can be also obtained by searching through Google or Najdi.si. However, being able to analyse the complete downloaded corpus enables much more powerful information extraction methods to be used, potentially leading to abuse of personal data. This is why we make the corpus available for research only, and require a short explanation of the use it will be put to. However, we (will) make available the metadata of the corpus, in particular the list of URLs included in it, which enables other to make their own corpus on this basis.

4. Comparative corpus analysis

This section investigates how different the slWaC$_2$ corpus is from its predecessor, slWaC$_1$ and from the KRES balanced reference corpus of Slovene (Logar et al., 2012). For this we used the method of frequency profiling, introduced by (Rayson and Garside, 2000). We first made a frequency lexicon of the annotation under investigation (lemma or grammatical description) for slWaC$_2$ and the corpus it was compared with, and then for each item in this lexicon computed its log-likelihood (LL). The formula takes into account the two frequencies of the element as well as the sizes of the two corpora which are being compared; the greater LL is, the more the item is specific for one of the corpora. To illustrate, we give in Table 3 the first 15 lemmas with their LL score and their frequency per million words in slWaC$_1$ and slWaC$_2$, with the larger frequency in bold.

Lemma	LL	slWaC$_1$pm	slWaC$_2$pm
člen	30,366	0.131	**0.282**
foto	23,092	0.018	**0.081**
m2	22,826	0	**0.033**
biti	22,767	**76,984**	74,493
°	21,447	0.001	**0.036**
3d	17,738	0	**0.026**
spoštovan	11,177	0.019	**0.059**
2x	11,092	0	**0.016**
tožnik	9,909	0.008	**0.036**
odstotek	9,265	**0.515**	0.393
co2	9,090	0	**0.013**
amandma	8,992	0.007	**0.031**
hvala	8,954	0.106	**0.173**
1x	8,505	0	**0.012**
ekspr	8,373	0	**0.012**

Table 3: The first 15 lemmas with highest log-likelihood scores and their frequency per million words for the comparison of the old and new version of slWaC

As can be noted, most of these highest LL lemmas are more prominent in slWaC$_2$; only *"biti" (to be)* and *"odstotek" (percent)* are more frequent in slWaC$_1$. Furthermore, quite a few lemmas have frequency 0 in slWaC$_1$. This is indicative of a difference in annotation between the two corpora: as mentioned, the tokenisation module of ToTaLe had been somewhat improved lately, which is evidenced in the fact that strings, such as "m2" and "3d" were wrongly split into two tokens in slWaC$_1$ but are kept as one in slWaC$_2$. It is a characteristic of LL scores that they show such divergences, which should ideally be fixed, to arrive at uniform annotation of the resources.

4.1. Lemma comparison with slWaC

The motivation behind comparing the previous and current version of slWaC was primarily to investigate what kind of text types are better represented in the new (or old) version of the corpus. Apart from the already mentioned differences in tokenisation, slWaC$_2$ is more prominent in three types of lemmas (texts). First, there are legal texts, (characterised by lemmas such as *"člen" (article,)* *"odstavek" (paragraph)*, *"amandma" (amendment)* *"tožnik" (plaintiff)*), which come predominantly from governmental domains, e.g. for "člen" mostly from uradni-list.si (official gazette), dz-rs.si (parliament), sodisce.si (courts). Second are texts that address the reader (or, say, parliamentary speaker) directly (*"spoštovan" (honoured)*, *"pozdravljen" (hello)*. For "spoštovan", the most highly ranked domains are, again, the parliament, i.e. dz-rs.si, followed by vizita.si (medical help page of commercial POP.TV), delo.si (main Slovene daily newspaper), in the latter two mostly from user forums. The corpus is thus more representative in text-rich domains whose content changes rapidly and that contain user-generated content. Third, the list contains two interesting "lemmas" with very high LL scores. The first is "ekspr" (only 19 in slWaC$_1$ but more than 9,000 in slWaC$_2$), which is the (badly tokenised) abbreviation "ekspr." meaning "expressive". It turns out that practically the only domain that uses this abbreviation is bos.zrc-sazu.si, i.e. the portal serving the monolingual Slovene dictionary SSKJ, which was newly harvested in slWaC$_2$. Similarly, the word "ino" (less than 500 in slWaC$_1$ but more than 7,000 in slWaC$_2$) turns out to be the historical form of *"in" (and)*. Practically the only domain containing this word (6,000x) is nl.ijs.si, which now hosts a large library of old Slovene books. The new slWaC thus contains some extensive new types of texts coming from previously unharvested domains or domains that have had large amounts of new content added. Finally, it is worth mentioning that the first slWaC$_2$ proper noun appears only at position 36 in the LL list, and is "bratušek" with almost 6,000 occurrences, referring to Alenka Bratušek, the former PM of Slovenia.

It is also instructive to see which lemmas are now less specific against slWaC$_1$. Interestingly, the greatest drop in frequency concerns the auxiliary verb *"biti" (to be)*. As all texts contain this lemma, it is difficult to analyse where this difference comes from, but our hypothesis is that legal texts, of which there are now significantly more, are more likely to use the present tense and passive constructions, which are made without the auxiliary. Among function words, there are less particles *"pa" ("but")*, used more in informal texts and less of *"da" (that)*, used to introduce relative clauses. One verb is much less used, *"dejati" (say, formal register)*, indicating a drop in the proportion of news items, where reporting on what a certain person said is quite frequent. Most of the list of course consists of nouns: in slWaC$_2$ there is relatively less written about *"odstotek" (percent)*, *"delnica" (share)*, *"milijon" (million)*, *"premier" (prime minister)*, *"predsednik" (president)*, *"dolar" (dollar)*, *"zda" (USA)*, again indicating less news and also the shifting of major news topics. Also, *"evro" (Euro)* is used less, but then the Euro symbol is used a lot more.

4.2. Lemma comparison with KRES

With slWaC$_2$, as with Web corpora in general, it is an interesting question of how representative and balanced they are. The easiest approach towards an answer is a comparison with "traditional" reference corpora, and such experiments have been already performed, e.g. between the British Web corpus ukWaC and BNC, the British National Corpus (Baroni et al., 2009). The comparisons have shown that while web corpora are different from classical corpora, which contain mostly printed sources, the differences are in general not great and so they can function as modern-day reference corpora.

We made a comparison between slWaC$_2$ and KRES (Logar et al., 2012), which is the balanced reference corpus of Slovene with 100 million words, sampled from Gigafida, the representative corpus of contemporary Slovene. Gigafida (*ibid*) contains texts from 1990 to 2012. The comparison shows that, as with slWaC$_1$, some of the differences are due to the different linguistic analyses. As mentioned, slWaC$_2$ was processed with ToTaLe, while KRES used the Obeliks tokeniser, tagger and lemmatiser (Grčar et al., 2012), and the two disagree in some lemmatisations, the most prominent being *"veliko/več" (much)*, *"mogoče/mogoč" (possible)*, *"edini/edin" (only)*, *"desni/desen" (right)*, *"levi/lev" (left)*, *"volitve/volitev" (elections)*, as well as some differences in tokenisation, e.g. "le-ta" and "d.o.o." as one token or three.

Real linguistic differences concern mostly two types of lemmas. The first are highly ranked non-content words such as *"pa, tudi, ter, naš" (but, also, and, our)*, which most likely show the bias of slWaC to informal writing. The second are content lemmas, which fall into several groups: *"spleten" (Web)*, *"podjetje" (company)*, *"tekma, ekipa" (match, team)*, *"volitve" (elections)*, *"sistem, uporabnik, aplikacija" (system user, applications)*, and "blog", i.e. slWaC has more commercial, sports, political and computer related texts, and, of course, texts specific to the web (blogs).

Conversely, KRES shows more lemmas to do with legal texts, such as *"člen, odstavek, zakon" (article, paragraph, law)*, so that even with slWaC$_2$ having more texts of this type than slWaC$_1$, it still has much less than KRES. KRES also has many more of two highly specific lemmas: *"tolar"" (former Slovene currency)* shows that KRES is by now already dated, while "wallander", the hero of a series of detective novels, shows that KRES – at least in this instance – has too much text from a single source, in this case a book series.

4.3. Grammatical comparison with KRES

Apart from lemmas, it is also interesting to compare how the distribution of morphosyntactic categories of slWaC$_2$ differs from that of KRES. To this end we calculated six LL comparison scores, for uni-, bi- and tri-grams of part-of-speech (PoS) and of complete morphosyntactic descriptions (MSDs).

The uni-gram PoS LL scores show that slWaC has significantly more adjectives, unknown words, conjunctions,

prepositions and particles, in this order. However, it has much less punctuation and numerals, and slightly less interjections. Esp. with unknown words and punctuation the differences might be, at least partially, an artefact of different annotation programs. For the others, the results show that slWaC tends more towards informal, user generated language, although this conclusion is somewhat offset by the fact that it has less interjections. However, tagging interjections is notoriously imprecise, and the difference here might also be due to different taggers used. Conversely, KRES with its numerals shows a preponderance of newspaper texts, which tend to use lots of dates, times, amounts, and sports scores.

PoS bi-grams again highlight the different annotation tools used. The most prominent combination in slWaC is a numeral followed by an abbreviation, e.g. "90 EUR, 206 kW, 298,80 m2" but this difference is due to the fact that in slWaC "EUR", "kW" etc. are treated as abbreviations, whereas they are common nouns in KRES. The same reasoning applies to combinations with punctuation. However, there are also legitimate combinations in the top scoring LL PoS bi-grams: slWaC has more noun + verb, adjective + noun and verb + adjective combinations, while KRES has more numeral + numeral, numeral + noun and verb + verb combinations. Scores for PoS tri-grams give little new information: apart from annotation differences, the most prominent slWaC combination is noun + noun + verb, which are mostly name + surname + predicate, e.g. "Oto Pestner naredil", while the most prominent for KRES is a sequence of three numerals.

As for MSDs, the differences in unigrams in favour of slWaC_2 are greatest for the three unknown word types that KRES doesn't use (Xf: foreign word, Xp: program mistake and Xt: typo), followed by general adverbs in the positive degree, coordinating conjunctions, present tense first person auxiliary verb in the plural ("smo") and animate common masculine singular noun in the accusative, i.e. the object of a sentence, e.g. "otroka". Conversely, KRES has much more punctuation, digits, common masculine and feminine singular nouns in the nominative (i.e. subjects) and general adverbs in comparative and superlative degrees. Bigrams show that slWaC has many more general adjective + common noun combinations in various genders and cases, while KRES has many more combinations with digits. The space of MSD trigrams is very large, and, if we discount the combinations appearing as a result of different annotations, does not show very interesting differences.

5. Conclusion

The paper presented a new version of the Slovene Web corpus, which is almost three times larger than its initial version and is made available through a powerful and freely accessible concordancer. During the construction process we focused on the content reductions obtained through near-duplicate removal, showing that both reductions to document and paragraph level remove a similar amount of content. We also compared the content of the slWaC_2 corpus to the slWaC_1 corpus and to the reference corpus KRES via frequency profiling on lemmas and grammatical descriptions. This comparison showed that the new version of the corpus has significantly more legal texts and specific text types, such as a dictionary and a library of historical books and (comparatively) less news. In the lemma comparison with KRES it has less legal texts but more user generated content and more commercial, sports, political and computer related texts, while the comparison of grammatical categories also shows a bias to informal writing as well as against newspaper items. But maybe the most surprising (although, in retrospect, quite logical) insight of the comparison using frequency profiling is that it is a very good tool to detect even slight differences in the processing pipelines used for the compared corpora, which then lead to significant differences in the (token, lemma and MSD) vocabularies.

There are several directions that our future work could take. First, by constructing the second version of two out of four existing web corpora of South Slavic languages, two ideas have emerged: one is to build a multilingual corpus consisting of all South Slavic languages, and the second to develop a monitor corpus which would be automatically extended with new crawls in predefined time frames. The second direction is in the annotation of the corpus, where more effort should be invested in developing a gold standard processing pipeline, which could then be used to reannotate the Slovene corpora in a unified manner. In addition, given that the Web contains a significant portion of user generated content containing non-standard language, the annotation pipeline should be extended by introducing a standardisation (normalisation) step on word-forms, similar to our approach to modernisation of historical Slovene words (Scherrer and Erjavec, 2013), which would then give better lemmas and MSDs, allowing for easier exploration of Web corpora.

6. References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Tomaž Erjavec and Simon Krek. 2008. The JOS Morphosyntactically Tagged Corpus of Slovene. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Tomaž Erjavec, Camelia Ignat, Bruno Pouliquen, and Ralf Steinberger. 2005. Massive multilingual corpus compilation: Acquis Communautaire and ToTaLe. *Archives of Control Sciences*, 15(3):253–264.

Miha Grčar, Simon Krek, and Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. In *Zbornik Osme konference Jezikovne tehnologije*, Ljubljana. Jožef Stefan Institute.

Emiliano Guevara. 2010. NoWaC: A Large Web-based Corpus for Norwegian. In *Proceedings of the NAACL HLT 2010 Sixth Web As Corpus Workshop*, WAC-6 '10, pages 1–7.

Nikola Ljubešić and Antonio Toral. 2014. caWaC - a Web Corpus of Catalan and its Application to Language Mod-

eling and Machine Translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Nikola Ljubešić and Tomaž Erjavec. 2011. hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. In Ivan Habernal and Václav Matousek, editors, *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, Lecture Notes in Computer Science, pages 395–402. Springer.

Nikola Ljubešić. 2014. {bs,hr,sr}WaC: Web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the WAC-9 Workshop*.

Nataša Logar, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt, and Simon Krek. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Zbirka Sporazumevanje. Trojina, zavod za uporabno slovenistiko: Fakulteta za družbene vede, Ljubljana.

Paul Rayson and Roger Garside. 2000. Comparing Corpora Using Frequency Profiling. In *Proceedings of the Workshop on Comparing Corpora*, pages 1–6. Association for Computational Linguistics.

Pavel Rychlỳ. 2007. Manatee/bonito – a modular corpus manager. *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70.

Yves Scherrer and Tomaž Erjavec. 2013. Modernizing historical Slovene words with character-based SMT. In *BSNLP 2013 - 4th Biennial Workshop on Balto-Slavic Natural Language Processing*, Sofia, Bulgaria.

Serge Sharoff. 2010. Analysing Similarities and Differences between Corpora. In *Proceedings of the Seventh Conference on Language Technologies*, pages 5–11, Ljubljana. Jožef Stefan Institute.

Drahomíra Spoustová, Miroslav Spousta, and Pavel Pecina. 2010. Building a Web Corpus of Czech. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Vít Suchomel and Jan Pomikálek. 2012. Efficient Web Crawling for Large Text Corpora. In Serge Sharoff Adam Kilgarriff, editor, *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pages 39–43, Lyon.

JANES se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino

Darja Fišer,* Tomaž Erjavec,† Ana Zwitter Vitez, *‡ Nikola Ljubešić₩†

* Oddelek za prevajalstvo, Filozofska Fakulteta
Aškerčeva 2, 1000 Ljubljana
darja.fiser@ff.uni-lj.si
† Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si
‡ Trojina, Zavod za uporabno slovenistiko
Dunajska 116, 1000 Ljubljana
ana.zwitter@guest.arnes.si
₩ Odsek za informacijske znanosti, Fakulteta za humanistične in družbene vede, Univerza v Zagrebu
Ivana Lučića 3, HR-10000 Zagreb
nikola.ljubesic@ffzg.hr

Povzetek

V prispevku predstavljamo vire, orodja in metodologijo, ki jih razvijamo za analizo nestandardne pisne spletne slovenščine. Ti so nujni za izdelavo sodobnih leksikografskih, normativnih in pedagoških priročnikov, ki brez podatkov o dejanski jezikovni rabi ni mogoča. Jezikovne modele, ki so dovolj robustni za obdelavo nestandardne pisne slovenščine, potrebujemo tudi za procesiranje spletnih besedil. Opisujemo gradnjo obsežnega korpusa pisne spletne slovenščine, izdelavo slovarja nestandardnih besed, tipičnih za pisno spletno komunikacijo, vrsto jezikoslovnih raziskav in razvoj metod za izboljšanje avtomatskega procesiranja nestandardne pisne spletne slovenščine. Razviti jezikovni viri bodo, primerno anonimizirani, ponujeni v odprt dostop pod licenco Creative Commons. Tako bodo omogočili prenos znanj na vsa področja, ki uporabljajo spletne vsebine, ki jih ustvarjajo uporabniki.

The JANES Project: methods, tools and resources for nonstandard Slovene

The paper presents an infrastructure and methodology under development for the analysis of user-generated content written in non-standard Slovene. They are indispensible in contemporary lexicographic, normative and pedagogic work, which cannot be comprehensive without information about real language use. Robust language models that can deal with nonstandard written Slovene are also needed for automatic text processing. A large and representative corpus of publicly available user-generated content and a web dictionary of non-standard Slovene will be compiled, comprehensive linguistic analyses will be performed and methods for automatic processing of non-standard text will be developed. The developed resources will be suitably anonymised and made openly available for download under the Creative Commons license. The developed resources, tools and methods will thus enable the transfer of knowledge to R&D in language technologies, lexicographic work and linguistic research.

1. Uvod

V času, ko računalniško posredovana komunikacija (ang. *computer-mediated communication*) in količina spletnih vsebin, ki jih na blogih in družbenih omrežjih ustvarjajo uporabniki, tako strmo naraščata, da je 90% tovrstnih besedil nastalo samo v zadnjih dveh letih (IBM 2013), postajajo njihove vsebine vse pomembnejši vir človeškega znanja in mnenj. Posledično se je povečala potreba po poznavanju in razumevanju t.i. internetnega jezika (*netspeak*), v katerem so te vsebine ustvarjene. Pisno spletno komunikacijo določajo okoliščine, kot so (ne)interaktivnost, (a)sinhronost, fizična (ne)prisotnost sogovornika in drugi situacijski dejavniki (Noblia 1998). Bolj kot je izbrana oblika komuniciranja interaktivna, poteka v realnem času in ima na drugi strani prisotnega sogovornika, več prvin spontanega govorjenega jezika vsebuje, vključno s (za računalniško komunikacijo prilagojenimi) paralingvističnimi in prozodičnimi elementi (Crystal 2001).

Za jezik pisne spletne komunikacije je značilna pogosta raba nestandardnih jezikovnih oblik, kot je nestandarden (bolj fonetičen) zapis besed, (npr. izključno male tiskane črke, opuščanje večine ločil in večkratno ponavljanje črk za čustveno poudarjanje zapisane izjave), in pogoste specifične okrajšave. Zaradi tega je jezikoslovna analiza in posledično tudi avtomatska obdelava tovrstnih vsebin otežena (Sproat idr. 2001), prizadevanja za premostitev teh ovir pa so trenutno ena bolj vročih tem na področju računalniškega jezikoslovja.

V sodobnem jezikoslovju so paradigme, ki na rabo nestandardnih jezikovnih različic v internetni pisni komunikaciji gledajo kot na odraz nepopolnosti ali osiromašenosti komunikacijskih zmožnosti, preživete, saj številne analize jezikovne rabe na internetu demonstrirajo sposobnost uporabnikov, da se prilagodijo računalniškemu mediju oziroma da zmožnosti medija izrabijo za zadovoljevanje svojih komunikacijskih potreb (glej npr. Tagg 2012), da si prizadevajo skrajšati in poenostaviti pisanje, predvsem pa da pisanje približajo svoji identiteti in govoru (Herring 2001). Večkrat je bilo dokazano tudi, da izpostavljenost nestandardnemu jeziku in njegova pogosta raba ne zmanjšujeta jezikovne zmožnosti (npr. Baron 2010).

Razkoraka med živostjo jezika in statičnostjo njegovega opisa ter iz tega izhajajočo nujno potrebo po raziskavah nestandardnega jezika se zavedajo tudi nekateri vodilni slovenski jezikoslovci, ki so že proučevali jezik SMS sporočil, spletnih forumov in elektronske pošte (npr. Logar 2003, Kalin Golob 2008, Dobrovoljc 2008, Jakop 2008, Michelizza 2008), kljub vsemu pa so tovrstne študije pri nas še vedno na obrobju interesa jezikoslovcev, zaradi česar je slovenski jezikoslovni prostor s tovrstnimi raziskavami izrazito podhranjen.

Zaenkrat se tudi najsodobnejša jezikovnotehnološka orodja, ki jih uporabljamo za procesiranje besedil, zelo slabo spopadajo z elementi nestandardnega jezika. Z njim imajo težave že povsem temeljna orodja, kot so na primer oblikoslovni označevalniki. Stanford tagger, eden najboljših označevalnikov na svetu, na standardnih angleških besedilih dosega 97 % natančnost, pri označevanju tvitov pa le 85 % (Gimpel idr. 2011).

Zato je cilj predstavljenih raziskav zapolniti eno največjih vrzeli slovenskega jezikoslovja: pomanjkanje virov, orodij in metodologij za jezik, ki se vedno bolj uporablja v vsakodnevni pisni komunikaciji in ki ga ustvarjajo vsi govorci slovenščine, ne zgolj novinarji, prevajalci, pisatelji ipd.

Tudi razvoj računalniškega jezikoslovja je odvisen od dostopnosti jezikovnih virov in orodij za obdelavo nestandardnega jezika. Pričakovani rezultati presegajo znanstveno relevantnost, saj bodo omogočili tudi razvoj najrazličnejših spletnih servisov in mobilnih aplikacij za slovenščino. Ti bodo imeli neposreden vpliv na zmanjševanje e-izključenosti govorcev slovenščine, ki trenutno iz pragmatičnih razlogov posegajo po tujejezičnih spletnih in mobilnih aplikacijah.

V nadaljevanju prispevka predstavljamo vire, ki jih bomo zgradili (razdelek 2), korpusnojezikoslovne analize, ki jih bomo opravili (razdelek 3), in orodja za računalniško obdelavo spletnih besedil, ki jih bomo razvili (razdelek 4). Prispevek sklenemo z razmislekom o razsežnostih in pomenu rezultatov raziskav za slovensko jezikoslovje in družbo.

2. Razvoj virov za proučevanje nestandardne pisne spletne slovenščine

Zgradili bomo reprezentativen korpus spletnih besedil, tipično zapisanih v nestandardnem jeziku, ki bo vseboval vsaj 20 milijonov pojavnic. Zajem besedil bo potekal avtomatsko, za kar bomo razvili namenska orodja. Osredotočili se bomo na javno objavljene pisne spletne vsebine in besedilne vrste, ki so tako po količini kot vplivu med najpomembnejšimi predstavniki nestandardnega jezika in zato najbolj relevantni za jezikoslovne raziskave. V korpus bodo vključeni:

- tviti (50 %)
- blogi (30 %)
- sporočila na forumih (10 %)
- komentarji na novice (5 %)
- komentarji na slovenski Wikipediji (5 %)

Z metodo za izbor bomo skušali zaobjeti čim bolj realno podobo tega dela slovenskega svetovnega spleta, da bo izdelan korpus zanj reprezentativen. Korpus bo vseboval natančne oznake besedilnih zvrsti, zato jih bo glede na konkretne raziskovalne potrebe mogoče proučevati tudi individualno in jih primerjati med seboj ter z drugimi korpusi.

2.1 Zajem besedil

Pri zajemu tvitov bomo razvili metodo, s katero bomo identificirali čim več slovenskih uporabnikov in sledili njihovo besedilno produkcijo. To bomo izvedli v naslednjih korakih:

- izdelava seznama visokofrekventnih polnopomenskih slovenskih besed, ki se ne pojavljajo v drugih jezikih,
- zajem množice tvitov s slovenskimi besedami,
- identifikacija dodatnih avtorjev in njihovih tvitov s pomočjo seznama sledilcev,
- razvoj natančnejših metod za identifikacijo slovenščine in izločanje tujih jezikov.

Za zajem blogov, forumov in komentarjev bomo nadgradili metodo gradnje splošnega spletnega korpusa slovenskih besedil (Ljubešić in Erjavec 2011). Fokusirano bomo pajkali samo domene, ki so bogate s temi tremi zvrstmi besedil. Pajkanje bo upoštevalo ime domene z ročno izdelanim seznamom bolj znanih jezikovnozvrstno specifičnih domen in s pomočjo spremljanja agregatorjev slovenskih blogov.

Enciklopedija Wikipedija je odprtodostopna, tako da je mogoče prevzeti celotno bazo, kar bo močno olajšalo identifikacijo komentarjev na posamezne strani in njihovo nadaljnje procesiranje.

2.2 Obdelava besedil

Avtomatsko zajeta besedila s spleta vsebujejo precejšnjo mero šuma (tudi do 80 %), kot je posredovanje nejezikovnih sporočil (fotografije, hiperpovezave ipd.), ki ga je treba odstraniti, da dobimo uporaben korpus besedil. Viri nestandardnega pisnega jezika na spletu pogosto vsebujejo mešanico slovenskih in tujih besed in črk, besedila so velikokrat napisana brez uporabe šumnikov, predvsem pa so posamezna besedila lahko zelo kratka, kar vse oteži delo programom za detekcijo jezika in njegovo označevanje.

Uporabnost korpusa je mnogo večja, če so besedila v njem jezikoslovno označena. Za slovenščino so bila zaenkrat razvita predvsem orodja za označevanje standardnega jezika, in sicer ToTaLe (Erjavec idr. 2005) in Obeliks (Grčar idr. 2012) za oblikoskladenjsko označevanje in lematizacijo, program DependencyParser (Dobrovoljc 2012) za skladenjsko analizo ter slNER (Štajner idr. 2013) in StandfordNER s slovenskim modelom (Ljubešić idr. 2013) za prepoznavanje imenskih entitet. Vendar predvidevamo, da bodo za obdelavo nestandardne pisne spletne slovenščine potrebne številne prilagoditve. Zato bomo avtomatsko zajeta besedila obdelali v naslednjih korakih:

1. čiščenje in deduplikacija spletnih vsebin (Ljubešić in Erjavec 2011),
2. identifikacija jezika prek seznama visokofrekventnih polnopomenskih slovenskih besed,
3. identifikacija preklapljanja med različnimi jeziki s statističnimi metodami,
4. identifikacija in poenotenje metapodatkov,
5. pretvorba v format XML po priporočilih TEI P5,
6. jezikoslovno označevanje (tokenizacija, oblikoskladenjsko označevanje, lematizacija, identifikacija imenskih entitet).

V tej fazi bomo z ročno evalvacijo identificirali najpogostejše napake obstoječih orodij za obdelavo standardnega pisnega jezika. Te napake bomo kasneje odpravili z izdelavo leksikona najbolj pogostih nestandardnih besed.

2.3 Izdelava spremljevalnega korpusa

Za splet je značilna velika dinamika produciranja besedil in hitro spreminjajoče se besedišče. Zato bomo vzpostavili prototipni sistem, ki bo sproti zajemal nove vsebine, jih občasno pretvoril, označil, indeksiral in ponudil v uporabo skozi konkordančnik. S tem bomo vzpostavili prvi slovenski spremljevalni korpus, ki bo omogočal sprotno spremljanje pisne spletne slovenščine ter zaznavanje novosti in sprememb na ravni leksike (neologizmi, naraščanje in upadanje rabe, vključevanje tujejezičnih prvin).

Taki postopki so že vključeni v delovni proces pri najsodobnejših leksikografskih projektih v tujini (Atkins idr. 2010), zanimivi pa so tudi za druge raziskave, ki se nanašajo na (ne)ustaljenost variant zapisa besed skozi čas, prilagajanje sloga in registra uporabnikov, spreminjanje diskurzivnih praks ipd. To je pri tako mladem in hitro razvijajočem se mediju zelo pomembno, saj se po eni strani uporabniki šele privajajo nanj, po drugi pa z razvojem tehnologije medij ponuja vedno nove funkcionalnosti, ki vplivajo tudi na rabo jezika.

3. Korpusna analiza nestandardne pisne spletne slovenščine

Poleg gradnje virov in orodij za nestandardno slovenščino je zelo pomembno osvetliti tudi rabo pisnega jezika na spletu iz različnih zornih kotov. Posebej se bomo posvetili sedmim jezikoslovnih raziskavam, ki bodo vsaka s svojega zornega kota osvetlile rabo pisne slovenščine na spletu. Rezultati raziskav bodo neposredno uporabni že pri razvoju orodij za računalniško obdelavo besedil, koristen pa bo tudi za številne druge jezikoslovne raziskave in jezikovnotehnološke aplikacije.

3.1 Primerjalna raziskava s pisnim standardom

Zbran in označen korpus spletne slovenščine bomo primerjali z referenčnim in uravnoteženim korpusom sodobnega slovenskega jezika KRES (Logar Berginc idr. 2012) s 100 milijoni pojavnic. Zasnovo za analizo smo na korpusu tvitov pripravili že v pilotni študiji (Erjavec in Fišer 2013), ki jo bomo sedaj razširili na celoten obseg korpusa JANES, ki poleg posodobljenega nabora tvitov vsebuje tudi štiri druge pomembne spletne besedilne zvrsti. V analizi bomo preučili:

- posebnosti zapisa (vzporedna raba večjega števila različic zapisa iste besede, npr. *itak/itaq, lahko noč/ln, počitniceeee*),
- leksikalne značilnosti (z metodo frekvenčnega profila (Rayson in Garside 2000), ki omogoča zaznavanje neologizmov in besedišča, ki je najbolj specifično za enega od korpusov),
- skladenjske značilnosti (kompleksnost povedi, besedni red in raba sklonov).

3.2 Primerjalna raziskava z govorom

Besedila tvitov, forumov in komentarjev nastajajo v okoliščinah, močno podobnih tistim, ki zaznamujejo govorno jezikovno produkcijo: avtor besedilo formulira kot neposreden odziv na družbeno dogajanje znotraj tesnih časovnih omejitev, poleg tega pa s strani naslovnikov pričakuje neposreden odziv na svoje jezikovno udejstvovanje. Govorne prvine nestandardne slovenščine bomo raziskali v primerjavi s korpusom govorjene slovenščine Gos (Verdonik, Zwitter Vitez 2011), pri tem pa bomo pozorni na:

- oblikoslovne posebnosti (primerjava deležev posameznih besednih vrst in analiza različic, ki se uporabljajo za eno standardno obliko),
- skladenjsko kompleksnost govornih in pisnih enot, stalne besedne zveze, besedni red in rabo sklonov,
- leksikalne specifike posameznih spletnih in govorjenih zvrsti,
- značilnosti avtorjev različnih profilov na podlagi označenih metapodatkov korpusa Gos, ki vsebuje podatke o spolu, starosti, izobrazbi in geografski pripadnosti.

3.3 Kolokacije v nestandardni pisni spletni slovenščini

Do razlik med standardno in nestandardno jezikovno rabo pogosto prihaja tudi na ravni kolokacij. Tovrstna razhajanja (npr. *Kaj dogaja! / Ful dogaja! / Tebi pa dogaja!* ipd.) so za jezikoslovje pomembna, ker je vezljivost v jeziku navadno relativno stabilna in zato spremembe v tem segmentu nakazujejo smer razvoja jezika. Če se dovolj ustalijo, sčasoma lahko postanejo del standarda (npr. *rabiti* v pomenu *potrebovati: Ne rabiš nobene dodatne opreme.*). Luščenje kolokacij bomo izvedli z naslednjima postopkoma:

- identifikacija nadpovprečno pogoste sopojavitve besed glede na njihovo siceršnjo frekvenco v korpusu v orodju Sketch Engine (Kilgarriff idr. 2010) na podlagi vnaprej pripravljenih leksikogramatičnih vzorcev za slovenščino (Krek in Kilgarriff 2006),
- analiza napak pri ekstrakciji kolokacij in prilagoditev orodja CollTerm, ki smo ga razvili v prejšnjih raziskavah (Pinnis idr. 2012).

3.4 Terminologija v nestandardni pisni spletni slovenščini

Številni blogi in forumi obravnavajo zelo specifično tematiko (npr. *medicina*), zato bo v njih pogosta tudi raba terminologije. Podobno velja za določene komentarje o urejanju specifičnih gesel na Wikipediji in tematsko specifične uporabniške račune na Twitterju. Ker gre v številnih primerih za neformalni sporočanjski položaj, lahko pričakujemo, da se bo raba terminologije razlikovala od tiste, ki je uporabljena v bolj formalnih registrih. Zato nas bosta pri analizi zanimali raba terminoloških dvojnic (npr. *slikovna pika-piksl*) in stopnja razhajanja nestandardne terminologije od standardne (npr. *HTML format* namesto *format HTML*).

S temeljito analizo terminologije nestandardnega pisnega jezika na spletu bomo izboljšali avtomatski luščilnik terminov LUIZ (Vintar 2010) in ga prilagodili tudi za luščenje terminologije s spletnih besedil za tri različne izbrane domene: medicino, računalništvo in gastronomijo.

3.5 Analiza pomenskih premikov v nestandardni pisni slovenščini

V jeziku se stalno razvija in spreminja tudi pomen že obstoječih besed. Detekcija novih pomenov je velik in pomemben izziv za leksikografijo in posodabljanje slovarskih gesel. Spletne publikacije, blogi in družbena omrežja pa so zaradi množične priljubljenosti in živahne jezikovne rabe idealen vir tovrstnih informacij. Aktualen popis semantičnega inventarja potrebujejo tudi različne jezikovnotehnološke aplikacije, kot sta npr. odgovarjanje na vprašanja in strojno prevajanje.

V ta namen bomo razvili algoritem, ki na podlagi vnaprej določenih pomenov iz semantičnega leksikona sloWNet (Fišer idr. 2012) v skladu z načeli distribucijske semantike v korpusu detektira tiste pojavitve določene besede, ki glede na sobesedilo ni dovolj podobna nobenemu od že obstoječih pomenov v sloWNetu. Seznam kandidatov bomo nato ročno pregledali in z morebitnimi novimi zaznanimi pomeni sloWNet tudi razširili.

3.6 Prepoznavanje žaljivega govora na spletu

V korpusu spletnih besedil bomo identificirali elemente žaljivega govora, saj anonimnost "omogoča posameznikom, da se obnašajo na načine, ki so zelo različni od njihovega vsakdanjega predstavljanja v vsakdanjem svetu" (Praprotnik 2003). Rezultati raziskave bodo zanimivi za institucije, odgovorne za zagotavljanje kulture dialoga (varuh človekovih pravic, spletni portali novinarskih hiš, družbena omrežja ipd.). Elemente žaljivega govora bomo identificirali v naslednjih korakih:

- izdelava manjšega učnega korpusa besedil, ki so jih uporabniki zaznali kot neprimerne, žaljive ali sovražne,
- označevanje eksplicitnih elementov žaljivega govora,
- identifikacija načel odklona žaljivih elementov od standarda (npr. *hebite se, čfeurji* ipd.),
- opredelitev značilk za avtomatsko zaznavanje potencialno žaljivih segmentov v celotnem korpusu spletnih besedil.

3.7 Izdelava slovarja nestandardne pisne spletne slovenščine

Poleg korpusa spletnih besedil in spremljevalnega korpusa bomo izdelali tudi leksikalno bazo, ki nam bo služila kot osnova za izdelavo spletnega slovarja. Ta pomemben jezikovni vir bo vseboval gesla, tipična in specifična za nestandardno jezikovno rabo v pisnih besedilih na spletu. Povezan bo tudi z drugimi viri, ki omogočajo uvid v lemo, obliko ali varianto skozi konkordančnik, druge spletne slovarje, kot je npr. SSKJ ali iskalnike najdi.si in Google.

Izdelan slovar bo uporaben za učitelje in učence, prevajalce ter zainteresirano javnost, pa tudi kot vir informacij o nestandardni leksiki za izdelavo novega slovarja slovenskega jezika.

4. Orodja za računalniško obdelavo

Priprava virov in prilagajanje orodij za avtomatsko obdelavo spletnih besedil bo temeljil na izkušnjah izdelave virov za starejši slovenski jezik (Erjavec 2012), kar mdr. predvideva cikličen pristop k izdelavi in izboljšavi orodij, v katerem ročno preverjeni podatki služijo za izboljšanje avtomatskega označevanje, to pa omogoča kvalitetnejšo osnovo za nadaljnje ročno označevanje.

4.1 Izdelava ročno označenega učnega podkorpusa

Za vse jezikovnotehnološke raziskave je zelo koristen ročno označeni korpus, saj služi kot učna množica za induktivno generiranje jezikoslovnih modelov, uporaben pa je tudi kot testna množica, na kateri je moč ovrednotiti kvaliteto razvitih avtomatskih postopkov za jezikoslovna označevanja. Tak korpus je koristen tudi za jezikoslovce, saj se na oznake v njem lahko bolj zanesejo kot na avtomatske.

Iz zajetega korpusa bomo po vnaprej določenih kriterijih za reprezentativnost in uravnoteženost vzorčili posamezna besedila oz. dele daljših besedil, dobljeni podkorpus avtomatsko označili in s tem dobili osnovo za izdelavo zlatega standarda, pri čemer je predvidena velikost korpusa 100.000 pojavnic. Vsaka besedna oblika v korpusu bo označena s svojo ustreznico in lemo iz standardnega jezika, z oblikoslovno oznako in predvidoma tudi s površinskoskladenjsko odvisnostno povezavo.

4.2 Izdelava učnega leksikona

Na podlagi primerjave besedišča iz referenčnega korpusa KRES in izdelanega korpusa bomo izdelali učni leksikon nestandardne pisne spletne slovenščine, ki bo vsebovala vsaj 1.000 gesel in 10.000 besednih oblik. Leksikon bo vseboval gesla, definirana s standardno zapisano lemo (osnovno obliko), besedno vrsto in, kjer knjižni jezik nima ustreznice, najbližje knjižne sinonime ali razlago. Geslo bo vsebovalo vse identificirane besedne oblike te leme in zglede iz korpusa, opremljene z metapodatki.

Tako kot korpus bo tudi leksikon zapisan v XML / TEI, kar pomeni, da ga bo možno povezati ali po želji vključiti v druge leksikalne vire, kot npr. v novi slovar

sodobnega slovenskega jezika. Takšnega leksikona tudi ni težko pretvoriti v leksikon za raznovrstne jezikovnotehnološke aplikacije.

4.3 Prilagajanje jezikoslovnega označevanja

Ker vsebujejo spletne vsebine, ki jih ustvarjajo uporabniki, jezik, ki se razlikuje od standardnega, je točnost označevanja standardnih jezikoslovnotehnoloških orodij tu bistveno slabša. To se je pokazalo tudi v naši preliminarni raziskavi označevanja tvitov (Erjavec in Fišer 2013), v kateri je bil delež napak v tokenizaciji, oblikoskladenjskem označevanju in lematizaciji občutno višji kot za standardno slovenščino, saj jih je bilo med pregledanimi 500 lemami 22 % napačnih, medtem ko je za korpus ccKRES bilo napačnih samo 4 %. Največ težav je bilo z lematizacijo pogovorno zapisanih besed (npr. *jst, js, nism*), s tokenizacijo emotikonov, ki jih program obravnava kot ločene pojavnice (npr. *:d, :p*) in označevanjem dvoumnih besed, kot so pridevniki in prislovi (npr. *oblačno*).

Naš cilj je prilagoditi obstoječe metode in tehnologije, da bodo sposobne obdelovati tudi nestandardni pisni jezik. Izboljšanje bomo dosegli na podlagi izdelanih jezikovnih virov (ročno označenega učnega podkorpusa in leksikona) in izsledkov ročne evalvacije avtomatsko pripisanih oznak.

Kot osnova nam bodo služile metode, ki smo jih že razvili za avtomatsko označevanje starejših besedil (Erjavec 2013), kjer po (prilagojeni) tokenizaciji posodobimo besedne pojavnice, nato pa nad tako normaliziranim besedilom uporabimo standardne modele za oblikoskladenjsko označevanje in lematizacijo. Posodabljanje besed je potekalo s pomočjo učnega leksikona, za neznane besede pa na podlagi ročno napisanih pravil transkripcije, ki se jih izvaja v formalizmu končnih avtomatov.

Vendar smo v novejših raziskavah (Scherer in Erjavec 2013) dosegli boljše rezultate z metodo transkripcije, ki temelji na statističnem strojnem prevajanju. Metoda, ki v primeru standardizacije besed kot enoto ne uporablja besed, temveč posamezne črke, se nauči modela preslikav iz parov *nestandardna beseda : standardna beseda*, ki jih bomo zajeli iz leksikona.

Nad besedili s standardizirano leksiko lahko uporabimo modele za standardno slovenščino in že s tem izboljšamo oblikoskladenjsko označevanje, lematizacijo in skladenjsko označevanje. Vendar ima trenutni pristop dve slabosti: posamezne besede standardizira neodvisno od konteksta, te pa so lahko dvoumne (npr. *jest*, v *jaz* oz. *jesti*), po drugi strani pa pri nestandardnem jeziku ne prihaja do razlik samo v leksiki, temveč tudi v skladnji. Zato bomo raziskali tudi druge, kompleksnejše metode označevanja, kjer označevalnike dodatno učimo tudi na ročno označenem podkorpusu ali pozamezne korake označevanja na različne načine povezujemo.

5 Zaključek

V prispevku smo predstavili metode, vire in orodja za analizo nestandardne pisne spletne slovenščine, ki jih razvijamo v okviru nacionalnega projekta JANES. Rezultati projekta bodo korpus pisne spletne slovenščine z več kot 20 milijoni pojavnic, slovar nestandardne spletne slovenščine, korpusno podprt jezikovni opis pisne spletne slovenščine na ortografski, oblikoslovni, leksikalni, pomenski in skladenjski ravni ter viri in metode za izboljšanje avtomatskega procesiranja nestandardne slovenščine.

Korpus in slovar bosta omogočila sodobnejšo in celovitejšo izdelavo empirično zasnovanih leksikografskih, normativnih in pedagoških priročnikov, s čimer želimo prispevati k dvigovanju samozavesti govorcev pri uporabi slovenščine. Metode za procesiranje nestandardne slovenščine bodo po eni strani olajšale prodor empiričnih raziskav v jezikoslovje, prav tako pa bodo olajšale postopke poizvedovanja po informacijah, rudarjenja po besedilih in povzemanja besedil, kar bo govorcem slovenščine omogočilo dostop do produktov, ki bolje podpirajo slovenski jezik.

Posredni doprinos raziskave predstavlja jezikovno neodvisna metodologija izgradnje virov in orodij, zaradi katere bodo pristopi neposredno uporabni tudi za sorodne jezike, kot so hrvaščina, srbščina in bosanščina, ki tovrstnih virov in orodij še nimajo razvitih.

Ob upoštevanju varovanja osebnih podatkov bodo izdelani viri ponujeni v odprt dostop pod licenco Creative Commons (CC BY-SA – Priznanje avtorstva, deljenje po enakimi pogoji). Izdelane vire in orodja bomo prenesli na slovensko raziskovalno infrastrukturo jeziko(slo)vnih podatkov in servisov za raziskave v humanistiki in družbenih vedah CLARIN.SI[1], kjer se bodo tudi trajno vzdrževali.

Literatura

B. T. S. Atkins, A. Kilgarriff, M. Rundell. 2010. Database of Analysed Texts of English (DANTE): the NEID database project. *Zbornik konference Euralex*.

N. S. Baron. 2010. *Always On: Language in an Online and Mobile World.* Oxford University Press.

D. Crystal. 2001. *Language and the Internet.* Cambridge University Press.

H. Dobrovoljc. 2008. Jezik v e-poštnih sporočilih in vprašanja sodobne normativistike. *Slovenščina med kulturami, Zbornik Slavističnega društva Slovenije 19*.

K. Dobrovoljc, S. Krek, J. Rupnik. 2012. Skladenjski razčlenjevalnik za slovenščino. *Zbornik Osme konference Jezikovne tehnologije,* str. 42-47.

T. Erjavec. 2013. Posodabljanje starejše slovenščine. *Uporabna informatika*, 21/4, str. 186-195.

T. Erjavec, C. Ignat, B. Pouliquen, R. Steinberger. 2005. Massive multi lingual corpus compilation : acquis communautaire and totale. *Archives of Control Sciences* 15, str. 529-540.

T. Erjavec. D. Fišer. 2013. Jezik slovenskih tvitov: korpusna raziskava. *Družbena funkcijskost jezika: (vidiki, merila, opredelitve). Obdobja 33*, str. 109-116.

[1] Common Language Resources and Technology Infrastructure: http://clarin.eu/

T. Erjavec. 2012. Jezikovni viri starejše slovenščine IMP : zbirka besedil, korpus, slovar. *Zbornik Osme konference Jezikovne tehnologije*, str. 52-56.

D. Fišer, T. Erjavec, J. Novak. 2012. sloWNet 3.0: development, extension and cleaning. *Zbornik konference Global Wordnet Conference*, str. 113-117.

K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, N. A. Smith. 2011. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments *Zbornik konference Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, str. 42–47.

M. Grčar S. Krek, K. Dobrovoljc. 2012. Obeliks: statistical morphosyntactic tagger and lemmatizer for Slovene. *Zbornik Osme konference Jezikovne tehnologije*, str. 42–47.

S. C. Herring. 2001. Computer-Mediated Discourse. *The Handbook of Discourse Analysis*. Oxford: Blackwell Publishers, str. 612–634.

IBM (2013) http://www.ibm.com/smarterplanet/us/en/business_analytics/article/it_business_intelligence.html [22.03.2014]

N. Jakop. 2008. Pravopis in spletni forumi – kva dogaja?. *Slovenščina med kulturami. Zbornik Slavističnega društva Slovenije 19*, str. 315–327.

M. Kalin Golob. 2008. SMS-sporočila treh generacij. *Slovenščina med kulturami. Zbornik slavističnega društva Slovenije 19*, str. 283–294.

A. Kilgarriff, S. Reddy, J. Pomikalek, P. V. S. Avinesh. 2010. A Corpus Factory for Many Languages. *Zbornik konference LREC'10*.

S. Krek, A. Kilgarriff. 2006. Slovene word sketches. *Zbornik konference Jezikovne tehnologije*.

N. Ljubešić, M. Stupar, T. Jurić, Ž. Agić. 2013. Combining available datasets for building named entity recognition models of Croatian and Slovene. *Jezikovne tehnologije, Slovenščina 2.0*, 1/2, str. 35–57

N. Ljubešić, T. Erjavec. 2011. hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. *Zbornik konference Text, Speech and Dialogue*.

N. Logar. 2003. Kratice in tvorjenke iz njih - aktualna poimenovalna možnost. *Współczesna polska i słoweńska sytuacja językowa*, str. 131–149.

N. Logar., M. Grčar, M. Brakus, T. Erjavec, Š. Arhar Holdt, S. Krek. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina.

M. Michelizza. 2008. Jezik SMS-jev in SMS-komunikacija. *Jezikoslovni zapiski* 14/1, str. 151–166.

M. V. Noblia. 1998. The Computer-Mediated Communication: A New Way of Understanding The Language. *Zbornik konference Internet Research and Information for Social Scientists*, str. 10–12.

M. Pinnis, N. Ljubešić, D. Ştefanescu, I. Skadina, M. Tadić, T. Gornostay. 2012. Term extraction, tagging, and mapping tools for under-resourced languages. *Zbornik konference Terminology and Knowledge Engineering*, str. 193–208.

T. Praprotnik. 2003. Pragmatični vidiki žaljive komunikacije v računalniško posredovani komunikaciji – multipla perspektiva. *Teorija in praksa*, 40/3, str. 515–540.

P. Rayson, R. Garside. 2000. Comparing corpora using frequency profiling. *Zbornik konference Comparing Corpora*, str. 1–6.

R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf, C. Richards. 2001. Normalization of non-standard words. *Computer Speech and Language*, 15(3), str. 287–333.

T. Štajner, T. Erjavec, S. Krek. 2013. Razpoznavanje imenskih entitet v slovenskem besedilu. *Jezikovne tehnologije, Slovenščina 2.0*, 1/2, str. 58-81.

C. Tagg. 2012. *Discourse of Text Messaging*. London: Continuum.

D. Verdonik, A. Zwitter Vitez. 2011. *Slovenski govorni korpus Gos*. Ljubljana: Trojina.

Š. Vintar. 2010. Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation. *Terminology* 16(2).

Using crowdsourcing in building a morphosyntactically annotated and lemmatized silver standard corpus of Croatian

Filip Klubička, Nikola Ljubešić

Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences
University of Zagreb
Ivana Lučića 3, HR-1000 Zagreb
{fklubick,nljubesi}@ffzg.hr

Abstract

This paper describes the creation of a morphosyntactically tagged and lemmatized silver standard corpus by using crowdsourcing. A data set containing 50.322 tokens compiled from the Croatian web corpus hrWaC was annotated using TreeTagger and HunPos taggers trained on the SETimes.HR corpus. Tokens that the tools annotated differently were passed on to the crowd. The crowd looked through contested nouns, verbs and adjectives, while experts checked and corrected those that the crowd decided were incorrect, along with the remaining parts of speech the two taggers did not agree on. The evaluation of the crowdsourcing yielded a single worker's accuracy to be ∼90%, and that of the majority answer of three workers to be ∼97%. While intrinsic evaluation of the resource by calculating accuracy of morphosyntactic tags showed an improvement of 8%, extrinsic evaluation of the corrected corpus on the task of morphosyntactic tagging produced an accuracy increase of little over 1%. The results point to the conclusion that the use of crowdsourcing in creating and improving language resources is indeed useful, but in the case of using the improved resource for enhancing morphosyntactic tagging, given the amount of already available gold corpus data, accuracy should be improved by developing a lexicon.

Uporaba mnoienja pri izdelavi oblikoskladenjsko oznaenega in lematiziranega korpusa hrvaine kot srebrnega standarda

V prispevku opišemo postopek izdelave oblikoskladenjsko označenega in lematiziranega korpusa hrvaščine z uporabo množičenja. Podatkovna množica, ki vsebuje 50.322 pojavnic, je bila vzorčena iz hrvaškega korpusa spletnih besedil hrWaC in oznaena z označevalnikoma TreeTagger in HunPos, ki sta se naučila modela jezika iz korpusa SETimes.HR. Pojavnice, ki sta jih programa označila različno, so bile z uporabo platforme za množičenje ffzgMnoštvo posredovane množici anotatorjev, ki so izmed obeh izbrali pravilno oznako. Množica je pregledala sporne samostalnike, glagole in pridevnike, medtem ko so eksperti pregledali in popravili tiste oznake, za katere se je množica odločila, da so napačne pri obeh označevalnikih, kot tudi preostale besedne vrste. Evalvacija množičenja je pokazala, da je natančnost posameznega anotatorja v povprečju ∼90%, večinska odločitev treh anotatorjev pa ∼97%. Medtem ko je intrinzina evalvacija vira z izraunom natančnosti oblikoskladenjskih oznak pokazala izboljšanje za 8%, je ekstrinzična evalvacija popravljenega korpusa pri nalogi oblikoskladenjskega označevanja povečala natančnost označevanja za malo več kot 1%. Rezultati kažejo, da je uporaba množičenja za izdelavo in izboljšanje jezikovnih virov koristna, vendar pa ne za izboljšanje oblikoskladenjskega označevanja, kjer bi bilo, glede na količino že dostopnih korpusnih podatkov kot zlatega standarda, moči bolje usmeriti v izdelavo leksikona.

Key words: crowdsourcing, silver standard, morphosyntactic annotation, lemmatization, Croatian language

1. Introduction

Crowdsourcing is a method that has lately been used more and more as a means for collecting data, as well as other kinds of organized effort in reaching certain goals. Whether it is crowdfunding, crowdvoting, crowdtagging or microworking,[1] the basic idea behind this method is that a vast number of people can contribute to a larger goal by doing little work individually. Due to its wide, interdisciplinary applicability, crowdsourcing is used more and more in the field of computer science for tagging data as a prerequisite for machine learning, a method applied in many fields, of which natural language processing is one.

The basic goal of this paper is to minimize the effort of building a large linguistic resource of acceptable quality, representative of the Croatian web. This silver standard corpus would be both lemmatized and morphosyntactically tagged. Though it need not improve any particular application, as simply creating a fresh linguistic resource is a worthwhile goal, we also evaluate the resource on certain natural language processing tasks.

The motivation behind including crowdsourcing into the procedure of building an annotated corpus is to simplify and speed up the process of checking and correcting the tags. The idea is that the crowd, whose work is time efficient and normally cheap or even, as in our case, free, can confirm which tags are correct. Consequentially, the expert, whose work is time consuming and, by comparison, expensive, needs to invest less time into checking the tags, focusing only on correcting those that are incorrect and problematic.

This approach represents a kind of middle ground between two approaches to tagging a corpus for machine learning - the classic approach, which is usually done so that an expert manually annotates raw data with no help (thus creating a so-called gold standard), and the more automated approach, where specialized tools automatically annotate the data, and the expert then later corrects the most probable mistakes (thus creating a silver standard). On the

[1] https://sites.google.com/site/crowdsourcewiki/

one hand, the problem is that manual annotation is time-consuming, tiresome and exhausting, but yields high accuracy rates, whereas on the other hand, tools for automatic tagging, though time-efficient, are imperfect and not precise enough. By putting crowdsourcing into the mix, the latter approach is enhanced, speeding up the expert's job.

The paper is structured as follows: after an overview of related work, follows a section that describes the workflow, covering sample selection and data preparation and our crowdsourcing tool. Section 4 describes the crowdsourcing and expert checks. In Section 5 we give intrinsic and extrinsic evaluation of the produced resource while we end with a conclusion in Section 6.

2. Overview of related work

As far as English is concerned, the problem of (statistical) morphosyntactic annotation is considered solved, as a very high per-token accuracy rate of 97.5% has been achieved (Søgaard, 2011), and though this is a recent development, it is not dramatically higher than the results reached by research in the decade preceding it. However, this is not the case for languages like Croatian, which are morphologically richer and have a looser sentence structure. The problem is actively being worked on and quite some progress has been made by following the statistical modeling paradigm: while earlier work (Agić et al., 2008b) achieved a 86.05% accuracy rate at the morphosyntactic level, but was not made available, the most recent work on the problem reaches 87.72% (Agić et al., 2013), resulting in the SETimes.HR corpus, an annotated corpus of Croatian language, which is publicly available[2], as are the models and test sets used in the paper.[3]

Alongside that, in another paper (Agić et al., 2010) the problem of MSD (morphosyntactic description) tagging is approached a bit differently, by using tagger voting, where the results of about a dozen automatic annotation tools are used as votes for the most likely morphosyntactic description, so that the answer given by the most taggers is considered correct. There is also the work of Peradin and Šnajder that approaches the problem from a different angle, by building rule-based grammars, which achieve an accuracy of 86.36%, but the systems are still in the prototype phase and not available as a ready-to-use tool (Peradin and Šnajder, 2012). Yet a third angle from which to approach the issue of lemmatization and MSD tagging is the approach of using a morphosyntactic lexicon during the annotation process(Agić et al., 2008a), but the result of this research is not publicly available.

Turning to languages related to Croatian, a few papers (Gesmundo and Samardžić, 2012b; Gesmundo and Samardžić, 2012a) dealt with lemmatization and tagging using a statistical approach. The models have been trained on the Serbian Multext East 1984 corpus and achieve an accuracy of 86.65% at the MSD level, but they are limited to the domain they were built on. Work has also been done in the past decade that provides an overview of a rule-based approach to the problem by utilizing NooJ and other similar tools.

However, when it specifically comes to using crowdsourcing for creating language resources and tools, and gathering linguistic data, aside from using it to clean up SloWNet[4] (Fišer and Tavčar, 2013), the Slovenian version of WordNet, such work has not been done on Croatian or other related languages. It is not very widespread when it comes to English either, especially if narrowed down to morphosyntactic tagging and lemmatization. There is a paper (Callison-Burch and Dredze, 2010) that provides an overview of the possibilities that Amazon's Mechanical Turk[5] offers in the field of (computational) linguistics, while there is also an effort to approach the crowdsourcing aspect of gathering linguistic data from a completely different angle – namely turning it into a game. Thus, Phrase Detectives[6] (Chamberlain et al., 2008) was designed, the first game created for collaborative tagging of language data on the internet.

3. Description of the workflow

3.1. Research outline

The basic corpus of text used in this research is a randomly selected sample of 5000 sentences from hrWaC 2.0, the second version of the Croatian Web corpus built from the .hr top-level domain, the construction of which is described in (Ljubešić and Klubička, 2014), and that encompasses some 1.9 billion tokens. Given that the idea is to build a high-quality linguistic resource for standard Croatian, but considering that sentences from the whole of the Croatian web vary in their quality and not all are standard Croatian sentences, the first step was to filter the sample. This task was delegated to the crowd and also served as a pilot testing of the crowdsourcing process, a detailed description of which is contained in section 4.

After the crowd completed the pilot task, the chosen standard sentences were annotated using two tools for automatic lemmatization and morphosyntactic tagging – TreeTagger[7] (Schmid, 1994; Schmid, 1995) and HunPos[8] (Halácsy et al., 2007). The tools were trained on the SETimes.HR corpus (Agić et al., 2013), a gold-annotated corpus of texts[9] collected from the Southeast European Times website[10] and the revised MULTEXT-East v4 morphosyntactic specifications that are used on the SETimes.HR corpus [11] were also used here as the annotation standard. The assumption was that those tokens that the tools annotated identically were tagged correctly, while those that the two taggers disagreed on were problematic. Out of 50322, there were 9965 such problematic tokens and they were passed on to the crowd for tagging in three phases – first the nouns, then the adjectives and finally the verbs. Based on the number of answers and the accuracy of the annotators, 3350 were declared correct.

[2] https://github.com/ffnlp/sethr
[3] http://nlp.ffzg.hr/resources/corpora/setimes-hr/
[4] lojze.lugos.si/darja/slownet.html
[5] https://www.mturk.com/
[6] http://anawiki.essex.ac.uk/phrasedetectives/index.php
[7] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[8] http://code.google.com/p/hunpos/
[9] http://nlp.ffzg.hr/corpora/setimes
[10] http://www.setimes.com/
[11] http://nlp.ffzg.hr/data/tagging/msd-hr.html

Afterwards, two experts looked over the lemmas and tags that the crowd had tagged as incorrect, as well as those of word classes the crowd did not annotate (such as adverbs, pronouns, prepositions, etc.) and corrected all those that needed correcting. This same approach was used in the creation of the Slovene jos1M corpus (Erjavec and Krek, 2008), only sans crodsourcing. Our corpus was also checked for non-existing tags and for non-agreement between adjectives and nouns, as well as between prepositions and following adjectives or nouns. With that, the silver standard corpus was completed and was then evaluated.

3.2. The ffzgMnoštvo crowdsourcing platform

The ffzgMnoštvo[12] crowdsourcing platform was used as a tagging tool to be used by the crowd. It is actually an adapted version of sloWCrowd (Fišer et al., 2014), adapted by Nikola Ljubešić for the purpose of this research.

Figure 1: ffzgMnoštvo user interface

After registering to the system, users can begin solving the task. They are offered a context that they need to judge, and the possible answers they can choose from are "Yes", "No" and "Don't know". Having registered to the system, users can also see how many answers they've given, as well as how many tasks are left in the database to be solved. To make things more interesting, a few gamification elements have been implemented into the platform, such as a progress bar and a hall of fame, which ranks users based on how much they contributed to the project. This is a way to add a healthy dose of competition between the annotators, which further motivates them to participate and solve tasks more regularly, while at the same time making the project more attractive (Chamberlain et al., 2008; Von Ahn, 2006).

4. Crowdsourcing linguistic data

Crowdsourcing via ffzgMnoštvo was done in four phases: 1. checking sentence standard and checking MSDs and lemmas of 2. nouns, 3. adjectives and 4. verbs.

4.1. Annotators

The annotators were exclusively students of the Faculty of Humanities and Social Sciences in Zagreb, attendees of the graduate course Selected Chapters from NLP at the Department of Information and Communication Sciences. The number of annotators varied from phase to phase, depending on how regularly they attended class. A quick demographics overview shows that most of them studied Informatics, Research Track, as a single major at the Department, and three of them had completed their undergraduate studies outside the Faculty. Those who studied a double major, along with Informatics, Research Track, also studied a philological program, be it Linguistics, English, Croatian, etc. Thus the annotators could initially be divided into two groups – those with a formal linguistic education and those without one. But it should be noted that the program at the Department significantly deals with language technologies, so even those who were only a single major actually had some of the required background knowledge. Possible discrepancies were made up for in the course itself, so all the annotators were adequately prepared for solving the task.

This kind of annotator demographic might be considered atypical of crowdsourcing, which usually includes several hundred, if not thousands of annotators, often with much more diverse backgrounds and expertise. However, this research was not intended to be large-scale crowdsourcing, but rather an experiment to measure how well a student group intrinsically motivated to take part in such a project, which is a feasible working force for the future, can solve such problems in a crowdsourcing environment.

4.2. Pilot phase – checking sentence standard

The data preparation phase contained a crowdsourcing pilot test, which, aside from filtering the initial sample, was done to try out the platform and familiarize the users with the system, concept and work principle, as well as to gain insight into how the crowd works and how to best utilize it in the following main phases.

Thirteen annotators were given 5000 sentences to annotate. The question they were asked was "Is the proposed sentence a standard Croatian sentence?" Each user annotated roughly 1000 questions, making up a total of 13167 answers.

One hundred sentences in the database were annotated ahead of time, representing a small gold standard set. The users sometimes got to annotate these golden sentences as well, not knowing the answer was predetermined. Their answers to the golden sentences served to calculate their accuracy and so provide feedback on their reliability – if they answered the gold standard sentences correctly, it can be concluded that they understand the task and that their answers to unannotated sentences are reliable. Calculating their accuracy was simply a matter of dividing the number of correctly answered gold standard sentences with the total number of gold standard sentences the users answered.

It is also important to see how many times a sentence has been tagged, as the goal is to gather as much data as possible by tagging as many sentences as many times possible. In an ideal scenario, each sentence would be tagged by at least two annotators.

The distribution shown in Table 1 shows that sentences were annotated quite a different number of times, i.e. that the variance of the distribution is quite high. This was due

[12] http://faust.ffzg.hr/ffzgmnostvo/

# of answers	# of sentences
0	285
1	941
2	1368
3	1246
4	698
5	462

Table 1: The distribution of the number of sentences by the obtained number of answers

to the algorithms of the ffzgMnotvo platform, and as it is obviously not the most economic way of collecting user responses, an additional intervention was made to the system by defining the number of maximum number of answers per task.

After the crowd tags the data, a final decision has to be made for each sentence on whether or not the crowd deems it standard or non-standard. The decision did not only take into account the distribution of answers, but also the user accuracy rates calculated for this task. So for each sentence that was answered more than 2 times, and there were 3774 such sentences, the accuracy values of the annotators that gave an answer were summed up in favor of that answer. For example, if two users, with accuracy rates of 0.72 and 0.65 said "Yes" to a sentence, and two, with accuracy rates of 0.88 and 0.86 said "No", then the final call for that sentence is "No" (not standard) because 0.88+0.86 $>$ 0.71+0.65. Thus, the crowd decided that 2831 sentences from the initial 5000 sentence set were standard, and these sentences made up the 50322 token corpus. The rest of the sentences were either judged as non-standard (1866 of them), tagged as "Don't know" (only 18), and due to the imperfections of the system, 285 sentences were not tagged even once.

We made an inquiry in the inter-annotator agreement between the users by calculating the Cohen's kappa (Berry and Mielke, 1988), which does not only take into account the observed agreement (Pr(a)), but also accounts for chance agreement (Pr(e)), as seen in equation 1.

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \quad (1)$$

The overall mean of IAA at the sentence standard task is 0.5169, while the mean of observed IAA is 0.7614, showing that there is quite some disagreement between annotators on the issue of sentence standard.

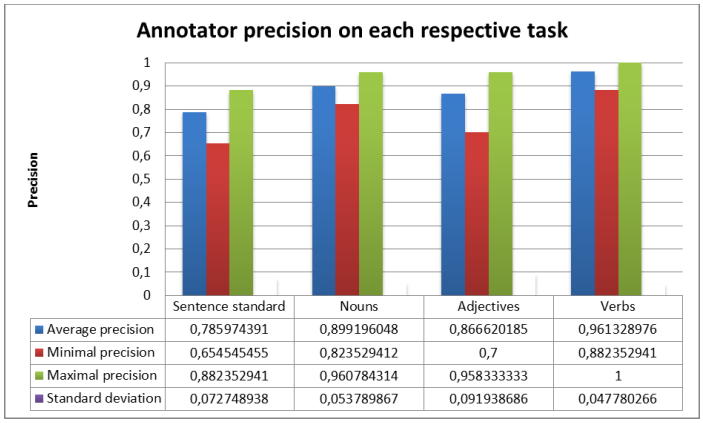
In the end, the sentence standard pilot testing confirmed that the system is applicable to the task, but it also showed that the crowd does not really agree on what a standard sentence is. It seems that whether a sentence is standard or not is open to interpretation and could create an issue when it comes to the quality of the final result. This could be prevented by giving very detailed instructions on what constitutes a standard sentence and by making the question completely unambiguous. Either that, or simply use crowdsourcing for gathering data on issues narrow enough to leave little to no room for interpretation.

4.3. Crowdsourcing MSD and lemmatization

The procedure as described in the former section, but adjusted in accordance with the insight gained from the pilot study, was repeated three more times on three new data sets: 14 annotators tagged 4896 nouns, 8 annotators tagged 2152 adjectives and 6 annotators tagged 478 verbs. The task was presented so that a context was given wherein the token of interest was marked in red, and the annotators were asked to judge the provided morphosyntactic description and the lemma of that token as correct, incorrect or unknown, as depicted earlier in Figure 1.

Of course, the question might arise of why the crowd did not do the whole annotation in the first place, but instead only judged the tags as correct or incorrect. We felt that the task for the crowd cannot be too complex, otherwise the feedback would be too slow and, probably, of low quality. Accordingly, we anticipated that there would not be much gain from delegating the difficult task of MSD tagging to the crowd workers, who are not experts, but rather decided to streamline to process. Furthermore, given the limits of the platform, coupled with the many grammatical categories in play, it would be near-impossible to implement and to properly adjust the interface.

A condensed annotator accuracy analysis shows that a single annotator's accuracy, on average, was about 90%, while that of the crowd collecting three answers was about 97%.



	Sentence standard	Nouns	Adjectives	Verbs
Average precision	0,785974391	0,899196048	0,866620185	0,961328976
Minimal precision	0,654545455	0,823529412	0,7	0,882352941
Maximal precision	0,882352941	0,960784314	0,958333333	1
Standard deviation	0,072748938	0,053789867	0,091938686	0,047780266

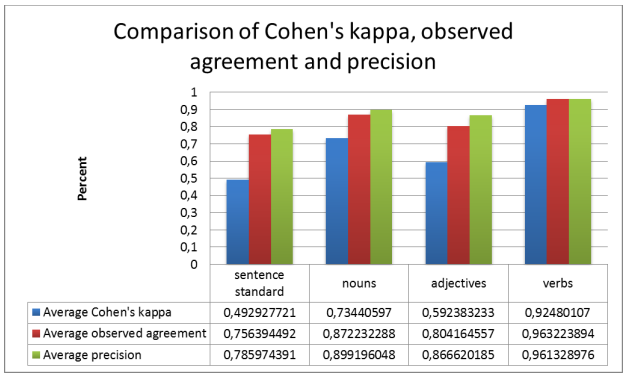
Figure 2: Annotator accuracy on each respective task

A calculation of the average IAA on the morphosyntactic disambiguation and lemmatization task (sentence standard was ignored due to the difference in the nature of the tasks) shows that the average kappa was about 75.05%, while the average observed agreement was 87.99%, as seen in Figure 3.

Annotators agreed much better when it came to MSDs and lemmas, because the task was a lot more unambiguous – if only one of the grammatical categories, or the lemma, was incorrect, the whole thing was to be declared incorrect.

4.4. Expert annotation

After the crowdsourcing was completed, the annotated corpus was split between two experts, whose task was to check and correct the tags that the crowd declared incorrect (2783 nouns, 1416 adjectives and 399 verbs), as well as



	sentence standard	nouns	adjectives	verbs
Average Cohen's kappa	0,492927721	0,73440597	0,592383233	0,92480107
Average observed agreement	0,756394492	0,872232288	0,804164557	0,963223894
Average precision	0,785974391	0,899196048	0,866620185	0,961328976

Figure 3: Accuracy and IAA comparison

the 2017 disputed tags of word classes the crowd did not annotate (pronouns, adverbs, prepositions, etc.) – making a total of 6615 tags to be checked and, if needed, corrected by the experts. These figures show a 42.12% reduction in the expert's workload thanks to the preceeding crowdsourcing step.

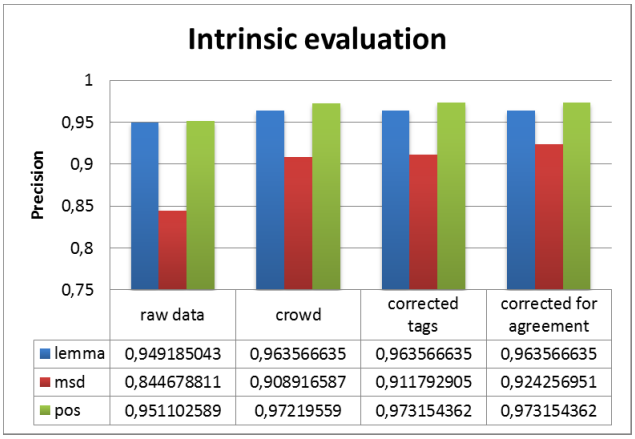
Following the experts' corrections of the tags, two additional steps were taken to further improve the tags – first, the corpus was checked for nonexisting tags (referred to as *corrected tags* in Figure 4), thus ruling out any typos or mistakes the experts might have made, as well as discrepancies and inconsistencies between the annotation standard in the SETimes corpus and the current data set. Second, the corpus was checked for non-agreement – noun phrases that had adjectives that did not agree in gender, number or case with the adjectives or nouns that follow them, as well as prepositions followed by adjectives or nouns that did not share their case (referred to as *corrected for agreement* in Figure 4).

5. Final resource evaluation

The result of the procedure is a lemmatized and morphosyntactically annotated corpus with a total of 50,322 tokens, which has been published and made publicly available on GitHub, along with the accompanying test sets described below.[13] To determine the quality of the final data, the corpus was evaluated on two levels – intrinsic and extrinsic. Intrinsic criteria are those connected to the goal of the system, whereas the extrinsic ones are connected to the system's function (Mollá and Hutchinson, 2003). So by doing an intrinsic evaluation of the corpus, the analysis would look at its accuracy in relation to itself – a sample from the corpus would be manually annotated, representing a gold standard, and it would be compared to that same segment taken from each phase of the corpus construction. Meanwhile, extrinsic evaluation analyzes the corpus' efficiency in a broader context of application, seeing how well it performs in use on some kind of NLP task. Such extrinsic evaluation can be done in at least two ways; either to use the corpus on its own as a resource for building a statistical tagging model, or to merge it with an already existing corpus and analyze its impact in a broader context, as an extension of already existing data for statistical modeling.

[13] https://github.com/ffnlp/sethr/

For the intrinsic evaluation, 50 sentences were chosen from the corpus of raw data. These sentences were annotated as a gold standard and were compared to the same sentences from every phase of the whole corpus annotation procedure. Three subtasks were taken into account – lemmatization, MSD tagging (providing the full grammatic description) and part-of-speech tagging (providing only the word class), and accuracy served as the evaluation metric. The evaluation showed that the accuracy of the corpus rose by 7.96% at the morphosyntactic level, 1.44% on the level of lemma and 2.2% on the part of speech (POS) level. A more detailed overview by each of the development phases can be seen in Figure 4 showing that the crowdsourcing and expert checking procedure produced most (∼80%) of the overall gain in accuracy.



	raw data	crowd	corrected tags	corrected for agreement
lemma	0,949185043	0,963566635	0,963566635	0,963566635
msd	0,844678811	0,908916587	0,911792905	0,924256951
pos	0,951102589	0,97219559	0,973154362	0,973154362

Figure 4: Results of intrinsic evaluation

The extrinsic evaluation was done in two rounds – first, statistical models were built using the HunPos(Halácsy et al., 2007) tagger, and this was done using data from the corpus in each of its development phases. The models were tried out on a separate test set of 6,429 tokens, or rather 300 sentences, of which 100 were taken from SETimes.HR, 100 from the Croatian Wikipedia and 100 from hrWaC.[14] Again, we used the same three tasks as during the intrinsic evaluation with accuracy as our evaluation metric. When comparing the models' initial and final accuracy on the test corpus, it rose by 0.4% at the part of speech level and 1.09% at the morphosyntactic level, but fell by 0.08% at the level of lemma.

	Raw data	Crowd	Final corrections
lemma	0.925	0.925	0.924
MSD	0.801	0.813	0.812
POS	0.952	0.957	0.952

Table 2: Results of standalone extrinsic evaluation

Second, the raw and final versions of the corpus were merged with an already existing annotated corpus, the SE-

[14] The first two data sets were built for (Agić et al., 2013), while web.hr.test was built for the purpose of this paper

Times.HR+[15] corpus and new models were trained on these data sets. Its accuracy on the aforementioned test set at the part of speech level rose by 0.1%, and by 0.17% at the morphosyntactic level, but fell by 0.73% on the level of lemma.

	SETimes+	Raw data and SETimes+	Final corrections and SETimes+
lemma	0.96	0.952	0.952
MSD	0.865	0.853	0.867
POS	0.97	0.969	0.971

Table 3: Results of extrinsic evaluation of expanded corpus

These results show that significant improvement can be achieved by using crowdsourcing for cleaning automatically annotated corpora, but that for the task at hand, given the amount of available gold standard data, minor or no improvement can be achieved.

6. Conclusion

The aim of this research has been fulfilled, as using crowdsourcing has shown itself to be a viable method for creating a silver standard dataset. This claim is backedup by the results of the evaluation performed on the resource. The intrinsic evaluation has shown a great rise in the accuracy of the data, so the positive effect of the crowdsourcing procedure is twofold – along with resulting with a high quality dataset, it also takes some of the weight off of the work the expert taggers do, making the procedure more economical.

The extrinsic evaluation is consistent in different environments – when merged with already existing corpora, the accuracy slightly grows at the MSD and POS levels, while it slightly falls at the level of lemma. These results suggest that the models have reached a plateau and that accuracy will not rise further if the quantity of training data is increased. At such a high accuracy level, it seems that there are so little inaccurate descriptions remaining that they become exceptions which statistical modeling alone cannot handle. Thus, the next logical step is to create a morphological lexicon and pair it with the annotation process, which would improve accuracy significantly.

Concerning adjusting the problem presentation on the crowdsourcing platform to the worker's perspective future work might deal with enhancing and speeding up the process by modifying the order of tasks – an error analysis can be done by looking at the differences between the tags in the initial and final stage of the corpus and then classifying the errors (whether the difference is only in the lemma/gender/case/a certain combination of categories). The tagging could thus be framed as solving groups of similar problems. Such an approach would take less cognitive effort from the annotators and would thus speed up the crowdsourcing.

[15]The SETimes.HR+ corpus is actually the SETimes.HR corpus (Agić et al., 2013) expanded with newspaper articles from various domains, amounting to a total of 135k tokens.

7. References

Ž. Agić, M. Tadić, and Z. Dovedan. 2008a. Combining part-of-speech tagger and inflectional lexicon for Croatian. *Proceedings of IS-LTC*.

Ž. Agić, M. Tadić, and Z. Dovedan. 2008b. Improving part-of-speech tagging accuracy for Croatian by morphological analysis. *Informatica*, 39(32):445–451.

Ž. Agić, M. Tadić, and Z. Dovedan. 2010. Tagger voting improves morphosyntactic tagging accuracy on Croatian texts. In *Information Technology Interfaces (ITI), 2010 32nd International Conference on*, pages 61–66. IEEE.

Ž. Agić, N. Ljubešić, and D. Merkler. 2013. Lemmatization and morphosyntactic tagging of Croatian and Serbian.

C. Callison-Burch and M. Dredze. 2010. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, page 112. Association for Computational Linguistics.

J. Chamberlain, M. Poesio, and U. Kruschwitz. 2008. Phrase detectives: A web-based collaborative annotation game. In *Proceedings of the i nternational conference on semantic systems (I-Semantics 08), Graz*.

T. Erjavec and S. Krek. 2008. The JOS morphosyntactically tagged corpus of Slovene. In *Proceedings of LREC*. Language Resources and Evaluation Conference.

D. Fišer and A. Tavčar. 2013. Več glav več ve: uporaba množičenja za čiščenje sloWNeta.

D. Fišer, A. Tavčar, and T. Erjavec. 2014. sloWCrowd: A crowdsourcing tool for lexicographic tasks. In *Proceedings of LREC*. Language Resources and Evaluation Conference.

A. Gesmundo and T. Samardžić. 2012a. Lemmatising Serbian as category tagging with bidirectional sequence classification. In *Proceedings of LREC*. Language Resources and Evaluation Conference.

A. Gesmundo and T. Samardžić. 2012b. Lemmatisation as a tagging task. In *Proceedings of ACL*. Association for Computational Linguistics.

P. Halácsy, A. Kornai, and C. Oravecz. 2007. HunPos: An open source trigram tagger. In *Proceedings of ACL*, pages 209–212. Association for Computational Linguistics.

N. Ljubešić and F. Klubička. 2014. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of EACL 2014*. Association for Computational Linguistics.

D. Mollá and B. Hutchinson. 2003. Intrinsic versus extrinsic evaluations of parsing systems. In *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: Are Evaluation Methods, Metrics and Resources Reusable?*, pages 43–50. Association for Computational Linguistics.

H. Peradin and J. Šnajder. 2012. Towards a constraint grammar based morphological tagger for Croatian. *Text, Speech and Dialogue*, 14:174–182.

H. Schmid. 1994. Probabilistic part-of-speech tagging us-

ing decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

H. Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*. Association for Computational Linguistics.

A. Søgaard. 2011. Semisupervised condensed nearest neighbor for part-of-speech tagging. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 48–52. Association for Computational Linguistics.

L. Von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.

Experiments with Neural Word Embeddings for Croatian

Leo Zuanović, Mladen Karan, Jan Šnajder

University of Zagreb, Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
{leo.zuanovic, mladen.karan, jan.snajder}@fer.hr

Abstract

Word representations extracted from a large corpus have been shown to be very useful in a variety of natural language processing tasks. Recently, there has been much work on using neural networks to learn good word representations from raw text. We adopt this approach and train neural word embeddings from a large Croatian web corpus. We evaluate the embeddings on three lexico-semantic tasks: synonym detection, semantic relatedness, and analogy modeling. Results on all three tasks are remarkably good and some of them markedly above the state-of-the-art results for Croatian. In particular, on the synonym detection and semantic relatedness tasks, the model achieves an accuracy of 73% and a correlation of 0.67 with human judgments, respectively.

Eksperimenti z nevronskimi vstavki besed za hrvaščino

Predstavitve besed, izluščene iz velikega korpusa, so se izkazale kot zelo koristne za raznovrstne naloge pri računalniški obravnavi naravnega jezika. V zadnjem času je bilo izvedenih veliko raziskav uporabe nevronskih mrež za učenje dobrih predstavitev besed iz neobdelanega besedila. V prispevku prevzamemo ta pristop in ga iz velikega korpusa hrvaščine naučimo nevronskih vstavkov besed. Vstavke evalviramo na treh leksikalnosemantičnih nalogah: detekciji sinonimov, semantični sorodnosti in modeliranju analogij. Rezultati na vseh treh nalogah so izredno dobri in nekateri bistveno boljši kot najboljši trenutni rezultati za hrvaščino. To še posebej velja za detekcijo sinonimov, kjer model doseže natančnost 73 %, ter za semantično sorodnost, ker model doseže korelacijo 0,67 s človeškimi odločitvami.

1. Introduction

In many natural language processing (NLP) tasks, model performance can be improved using word features induced from a large corpus, so-called *word representations*. A word representation is a mathematical object (typically a vector) associated with each word. *Distributional word representations* are derived from corpus-extracted co-occurrence matrix of words (rows) in some contexts (columns) (Turian et al., 2010). A number of design decisions have to be made when building such representations: the type and size of the context (e.g., a word window, a sentence, or document), how the counts are weighted (raw frequency, binary, tf-idf, etc.), and which dimensionality reduction technique to apply. Popular approaches include Latent Semantic Analysis (Deerwester et al., 1990), Random Indexing (Sahlgren, 2005), and Latent Dirichlet Allocation (Blei et al., 2003).

An alternative approach to word representations, which is gaining a lot of attention recently, is to learn a distributed representation in a supervised manner. Generally speaking, a *distributed representation* of a symbol is a vector of features, which characterize the meaning of the symbol while not being mutually exclusive, i.e., each of the features can be independently active or inactive, thus enabling the characterization of an exponential number of symbols (Bengio, 2008). In particular, distributed representations of words are called *word embeddings*, because the words are embedded into a dense, low-dimensional, real-valued vector space. The main idea is that functionally similar words will become close to each other after being embedded in this space.

In this paper, we experiment with word embeddings for Croatian using the recently proposed neural network-based models of Mikolov et al. (2013a). We evaluate these representations on three standard lexico-semantic tasks, namely synonym detection, semantic relatedness, and syntactic and semantic analogies. We show that the obtained word representations markedly outperform previous state-of-the-art results for Croatian.

2. Related work

Word embeddings are typically obtained as a by-product of training neural network-based language models (NNLMs). Language modeling is a classical NLP task of predicting the probability distribution over the "next" word, given some preceding words. In NNLMs, a sequence of words is first transformed into a sequence of word vectors via a projection matrix (weights between the input and the hidden layer), and then the network learns the probability distribution over these vectors. The advantage of using distributed representations is that they allow the model to generalize well to sequences that did not occur in the training set, but that are similar in terms of their features (i.e., their distributed representation), thus ameliorating the notorious data sparseness problem (Bengio, 2008).

NNLMs were first studied in the context of feed-forward networks (Bengio et al., 2003), and later in the context of recurrent neural network models (Mikolov et al., 2010; Mikolov et al.,). Computationally more efficient models were obtained by using hierarchical prediction (Morin and Bengio, 2005; Mnih and Kavukcuoglu, 2013a; Le et al., ; Mikolov et al., 2010; Mikolov et al.,).

Unlike the above-described architectures, which aim at learning good language models, the architectures described in (Mikolov et al., 2013a; Mikolov et al., 2013b) are primarily concerned with learning good word representations, and

therefore are free to move away from the paradigm of predicting the target word from the previous words. Since these are the models we used in this work, we describe them in more detail in the next section.

3. CBOW and continuous skip-gram models

In the Continuous Bag of Words (CBOW) model (Mikolov et al., 2013a), the training objective is to learn distributed representations of the surrounding words (both the preceding and the succeeding ones), which, when combined, are good at predicting the intermediate target word. In the continuous skip-gram model (Mikolov et al., 2013a), on the other hand, the objective is to learn a distributed representation of the input word that is good at predicting its context in the same sentence.

These neural architectures are perhaps more easily understood as a log-linear classifier. Given a sequence of training words $w_1, w_2, \ldots, w_T$, the objective of the skip-gram model is to maximize the average log-probability:

$$\frac{1}{T}\sum_{t=1}^{T}\left[\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)\right]$$

where c is the size of the training context (window).[1] That is, the correct labels for the current word w_t are its surrounding words, w_{t+j}. In contrast, the CBOW model aims to maximize the probability $p(w_t|w_{t+j})$, i.e., the correct label for the surrounding words w_{t+j} is the intermediate word w_t.[2]

These models have two sets of word representations: one for the "input" words (w_{t+j} in the CBOW model and w_t in the skip-gram model) and one for the "output" (target) words (i.e., the words being predicted: w_t in CBOW and w_{t+j} in skip-gram model). These "input" representations are the ones we actually use for the semantic modeling of words. The conditional probabilities $p(w_t|w_{t+j})$ and $p(w_{t+j}|w_t)$ are defined as:

$$p(w_O|w_I) = \frac{\exp\left({v'_{w_O}}^{\mathrm{T}} \cdot v_{w_I}\right)}{\sum_{w=1}^{W}\exp\left({v'_w}^{\mathrm{T}} \cdot v_{w_I}\right)}$$

where v_w and v'_w are the "input" and "output" vector representations of w, and W is the number of words in the vocabulary. However, this formulation is impractical because the cost of computing $\nabla \log p(w_O|w_I)$ is proportional to W, so a computationally efficient approximation of the full softmax – hierarchical softmax that uses a Huffman binary tree representation of the output layer – is used instead. Other methods used to speed-up the computation are Negative Sampling and subsampling of frequent words.

The models are trained by minimizing the negative log-likelihood using stochastic gradient descent. The gradient is computed using the well-known backpropagation rule (Rumelhart et al., 1988). Training can be performed on a large corpus in a short time (billions of words in hours). Mikolov et al. (2013a) have shown that skip-gram gives better word representations when the data is small, whereas the CBOW is faster and more suitable for larger datasets.

For details, refer to Mnih and Kavukcuoglu (2013b), who provide a good introduction to this type of models and describe a more general log-linear model.

4. Experimental setup

For training the CBOW and continuous skip-gram models, we used the publicly available `word2vec` implementation.[3] All models were trained on fhrWaC, a filtered version of Croatian web corpus described in (Ljubešić and Erjavec, 2011; Šnajder et al., 2013).[4] The corpus consists of 51M sentences and 1.2G tokens. All the words that occurred less than five times in the training data were discarded from the vocabulary, which resulted in a vocabulary of 1.4M words.

The parameters we varied are: the type of the model (CBOW or skip-gram), vector size, and the size of the context window. In what follows, we name the models to reflect their parameters (e.g., skip_100_5 is a skip-gram model with 100-dimensional vectors and a context window of at most five words). We used a hierarchical softmax in the output layer and subsampled frequent words with a threshold of 10^{-3}. The training times range from less than an hour for the CBOW model to several hours for the skip-gram model.

4.1. Task 1: Synonym detection

We evaluate the embeddings on a standard task from lexical semantics, namely synonym detection. We use the dataset created by Karan et al. (2012), with word choice questions for nouns, verbs, and adjectives (1000 questions each).[5] Each question consists of one target word with four synonym candidates, of which one is correct. The questions were extracted automatically from a machine readable dictionary of Croatian. For instance, *težak (husbandman, farmer): poljoprivrednik (agriculturalist, farmer), umjetnost (art), radijacija (radiation), bod (point)*. To make predictions, we compute pairwise cosine similarities of the target word vectors with the four candidates and predict the candidate(s) with maximum similarity.

We compare against the LSA-based synonym detection model of Karan et al. (2012), which uses 500 latent dimensions and paragraphs as contexts (LSA500P), and against a similar model that uses documents as context (LSA500D). We also compare against a Distributional Memory model of Šnajder et al. (2013), which is a state-of-the-art model on this task for Croatian.

4.2. Task 2: Semantic relatedness

For the semantic relatedness task, we use the dataset created by Janković et al. (2011), [6] containing 450 word pairs with human-annotated semantic relatedness judgments on a scale from 1 to 5. The annotations were made by 12 judges,

[1] The parameter c is actually the maximum window size. For each target word, a number R is drawn randomly from the $[1, c]$ range, and then R neighboring words to each side are taken.

[2] The surrounding words are not presented one-by-one, rather their vector representations are averaged and the resulting vector is used as the input to the classifier. We can consider this vector to be a predicted representation of the target (middle) word.

[3] `https://code.google.com/p/word2vec/`

[4] `http://takelab.fer.hr/data/fhrwac/`

[5] `http://takelab.fer.hr/data/crosyn/`

[6] `http://takelab.fer.hr/data/crosemrel450/`

out of which six with strongest agreement were selected and their scores averaged. For example, the pair *mlad (young) – star (old)* is assigned a score of 5.0, while the pair *utorak (Tuesday) – srijeda (Wednesday)* is assigned a score of 4.5.

As in the previous task, we use cosine as the similarity measure. We compare the computed similarities against the human judgments using Pearson's and Spearman's correlation coefficients. We use LSA500D as the baseline model.

4.3. Task 3: Syntactic and semantic analogies

Mimicking the experiments presented by Mikolov et al. (2013b), we also evaluate the embeddings on two analogy-based challenge sets. These consist of questions of the form "*a* is to *b* as *c* is to __", denoted as $a : b \rightarrow c : ?$. The task is to correctly predict the omitted fourth word, with only the exact word match deemed correct. Let $\mathbf{x}_a, \mathbf{x}_b, \mathbf{x}_c$ be the corresponding word embeddings (all normalized to unit norm). Then the expected answer to $a : b \rightarrow c : ?$ is given by $\mathbf{y} = \mathbf{x}_b - \mathbf{x}_a + \mathbf{x}_c$. Of course, there might not exist a word at that exact position in the vector space, thus we search for a word w^* that is most similar to word y (excluding the input question words):

$$w^* = \arg\max_w \frac{\mathbf{x}_w \mathbf{y}}{\|\mathbf{x}_w\| \|\mathbf{y}\|}$$

We test the syntactic analogies on the task of finding the correct comparative form of an adjective. To build the dataset, we first selected 50 adjectives with frequent comparatives in the corpus. Next, out of those 50 adjectives, we selected 10 most common adjectives (and their comparatives), and for each we randomly selected 35 out of the 49 remaining pairs. Then, each of the 10 most common pairs is written down 35 times followed by the corresponding 35 pairs, yielding a total of 350 questions. An example item is *bogat (rich)* : *bogatiji (richer)* → *opasan (dangerous)* : ? [*opasniji (more dangerous)*]. The motivation for how the dataset was constructed is that the ten most common pairs will very well capture the "idea" of the comparative form.

To test the semantic analogy, we use the set of most common countries and their capitals obtained by translating the English version of the dataset created by Mikolov et al. (2013a).[7] The form is similar to the comparatives set, with one of the 23 pairs being repeated 22 times, each time followed by a different pair, resulting in 506 (i.e., 22×23) questions. For example, *Tokio (Tokyo)* : *Japan* → *Pariz (Paris)* : ? [*Francuska (France)*]. We make the analogies dataset freely available.[8]

As a baseline, we use the LSA500D model. These vectors were learned over a lemmatized corpus, hence there are no vector representations for the comparative forms.

5. Results

Our preliminary experiments have shown that network parameters (sizes of the layers) influence the results considerably, especially in the semantic analogies task. In this work we did not perform a systematic parameter optimization and we leave this for future work. Nonetheless, it should be noted that in most cases, even with the worst parameter settings, the neural network models still outperformed the simpler models by a considerable margin.

Model	N	A	V
Dm.Hr	70.0	66.3	63.2
LSA500P	67.2	68.9	61.0
LSA500D	60.0	60.8	50.7
skip_100_5	71.9	69.9	71.3
skip_200_5	73.4	71.9	74.1
skip_200_10	75.6	72.6	70.1
skip_500_5	75.5	**73.0**	**75.8**
skip_1000_10	**76.8**	72.7	72.2
cbow_100_5	61.7	69.3	69.0
cbow_100_10	62.5	67.3	64.9
cbow_200_5	66.2	70.6	72.1
cbow_200_10	64.7	67.8	68.6
cbow_500_5	66.9	70.3	72.8
cbow_1000_5	66.6	70.3	72.1
cbow_1000_10	29.8	25.9	27.6

Table 1: Results for the synonym detection task.

5.1. Task 1: Synonym detection

Table 1 shows the results for the considered models on nouns (N), adjectives (A), and verbs (V). Word embeddings outperform the baseline models across all considered parts of speech. continuous skip-gram models generally perform better than CBOW models. Overall, the biggest improvement over the baselines is achieved for verbs and the smallest for adjectives. This could be due to the fact that Croatian adjectives can have more than 40 different forms, which results in over 40 word embeddings for a single word, while for the evaluation we only consider a single vector – that of the word's lemma. It would be interesting to investigate whether better word representations for lemmas could be obtained by averaging the vectors of all the different forms of a word or by training the models over a lemmatized corpus.

Regardless of parameter setting, the neural network models outperform the state-of-the-art synonym detection model (Dm.Hr) from Šnajder et al. (2013).

5.2. Task 2: Semantic relatedness

The results for the semantic relatedness task are given in Table 2. Word embeddings markedly outperform the baseline. Skip-gram models again outperform CBOW. Spearman's coefficient is lower than Pearson's, indicating the presence of outliers. We also conducted experiments on the version of the set where all 12 judges are included. This, expectedly, decreases the results slightly (by 1–2 points).

All neural network models substantially outperform the LSA baseline. We could not compare against the Random Indexing model from Janković et al. (2011), because the authors did not use correlation coefficients for evaluation.

5.3. Task 3: Syntactic and semantic analogies

The performance of various CBOW and skip-gram models on the word analogy set is shown in Table 3. We have no baseline for comparative forms of adjectives, but selecting

[7] Available from `http://goo.gl/OR5W05`

[8] `http://takelab.fer.hr/data/croanalogy`

Model	Pearson	Spearman
LSA500D	0.438	0.225
skip_100_5	0.670	0.575
skip_200_5	0.665	0.600
skip_200_10	**0.677**	0.591
skip_500_5	0.673	0.573
skip_1000_10	0.649	**0.623**
cbow_100_5	0.533	0.438
cbow_100_10	0.501	0.432
cbow_200_5	0.570	0.468
cbow_200_10	0.537	0.453
cbow_500_5	0.576	0.504
cbow_1000_5	0.560	0.490
cbow_1000_10	0.466	0.351

Table 2: Results for the semantic relatedness task.

Model	Comparatives	Capitals
LSA500D	–	30.9
skip_100_5	36.6	13.4
skip_200_5	47.1	18.6
skip_200_10	**48.3**	28.8
skip_500_5	42.0	24.9
skip_1000_10	34.0	**35.7**
cbow_100_5	30.3	9.3
cbow_100_10	24.6	9.1
cbow_200_5	31.4	8.4
cbow_200_10	28.9	9.1
cbow_500_5	31.1	10.8
cbow_1000_5	23.4	12.3
cbow_1000_10	0	0

Table 3: Results for the word analogy task.

the right word out of 1M words in almost 50% of cases is a remarkable result. Skip-gram models again outperform CBOW. The cbow_1000_10 performs suspiciously poorly; we believe this may be due to a technical issue in the training procedure (e.g., insufficient training iterations).

6. Conclusion

Distributed word representations (aka word embeddings) have gained a lot of attention recently. We have built word embeddings for 1.4M Croatian words using CBOW and continuous skip-gram models. We evaluated the embeddings on three lexico-semantic tasks, showing a remarkable improvement in performance over the state of the art for Croatian. The skip-gram model outperformed the CBOW model.

For future work, we intend to investigate how various preprocessing steps (e.g., lemmatization) and properties of the corpus influence word representations. Another line of research is the application of word embeddings to tasks such as POS tagging and named entity recognition.

7. References

Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Y. Bengio. 2008. Neural net language models. *Scholarpedia*, 3(1):3881.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. 1990. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

V. Janković, J. Šnajder, and B. D. Bašić. 2011. Random indexing distributional semantic models for Croatian language. In *Text, Speech and Dialogue*, pages 411–418. Springer.

M. Karan, J. Šnajder, and B. D. Bašić. 2012. Distributional semantics approach to detecting synonyms in Croatian language. *Information Society, Proc. of the Eighth Language Technologies Conference*, pages 111–116.

H.-S. Le, I. Oparin, A. Allauzen, J. Gauvain, and F. Yvon. Structured output layer neural network language model. In *Proc. of ICASSP'11*.

N. Ljubešić and T. Erjavec. 2011. hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. In *Text, Speech and Dialogue*, pages 395–402. Springer.

T. Mikolov, S. Kombrink, L. Burget, J. Cernocky, and S. Khudanpur. Extensions of recurrent neural network language model. In *Proc. of ICASSP'11*.

T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH*, pages 1045–1048.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

A. Mnih and K. Kavukcuoglu. 2013a. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in Neural Information Processing Systems 26*, pages 2265–2273.

A. Mnih and K. Kavukcuoglu. 2013b. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in Neural Information Processing Systems*, pages 2265–2273.

F. Morin and Y. Bengio. 2005. Hierarchical probabilistic neural network language model. In *AISTATS*, volume 5, pages 246–252.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. 1988. Learning representations by back-propagating errors. *Cognitive modeling*.

M. Sahlgren. 2005. An introduction to random indexing. In *Methods and applications of semantic indexing, TKE*, volume 5.

J. Šnajder, S. Padó, and Ž. Agić. 2013. Building and evaluating a distributional memory for Croatian. In *Proc. of ACL 2013*, pages 784–789.

J. Turian, L. Ratinov, and Y. Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of ACL 2010*, pages 384–394.

Automatic de-identification of protected health information

Jelena Jaćimović*†, Cvetana Krstev*, Drago Jelovac†

* University of Belgrade, Faculty of Philology
Studentski trg 3, 11000 Belgrade, Serbia
jjacimovic@rcub.bg.ac.rs
cvetana@matf.bg.ac.rs

†University of Belgrade, School of Dental Medicine
Dr. Subotića 8, 11000 Belgrade, Serbia
drago.jelovac@stomf.bg.ac.rs

Abstract

This paper presents an automatic de-identification system for Serbian, grounded on a rapid adaptation of the existing named entity recognition system. Based on a finite-state methodology and lexical resources, the system is designed to detect and replace all explicit personal protected health information present in the medical narrative texts, while still preserving all the relevant medical concepts. The results of a preliminary evaluation demonstrate the usefulness of this method both in preserving patient privacy and the de-identified document interoperability.

Avtomatska dezidentifikacija zaščitenih zdravstvenih podatkov

V prispevku predstavimo sistem za avtomatsko dezidentifikacijo v srbščini, ki temelji na hitri prilagoditvi obstoječega sistema za identifikacijo imenskih entitet. Sistem je zasnovan na metodologiji končnih avtomatov in jezikovnih virov ter identificira in zamenja vse eksplicitne zaščitene zdravstvene osebne podatke v medicinskih narativnih besedilih, pri čemer pa ohrani relevantne medicinske koncepte. Rezultati preliminarne evalvacije so pokazali uporabnost te metode, in sicer tako pri zaščiti osebnih podatkov pacientov kot pri interoperabilnosti dezindentificiranih dokumentov.

1. Introduction

Current advances in health information technology enable health care providers and organizations to automate most aspects of the patient care management, facilitating collection, storage and usage of patient information. Such information, stored in the form of electronic medical records (EMRs), represents accurate and comprehensive clinical data valuable as a vital resource for secondary uses such as quality improvement, research, and teaching. Besides the vast useful information, narrative clinical texts of the EMR also include many items of patient identifying information. For both ethical and legal reasons, when confidential clinical data are shared and used for research purposes, it is necessary to protect patient privacy and remove patient-specific identifiers through a process of the de-identification.

A de-identification is focused on detecting and removing/modifying all explicit personal Protected Health Information (PHI) present in the medical or other records, while still preserving all the medically relevant information about the patient. Various standards and regulations for health data protection define multiple directions to achieve the de-identification, but the most frequently referenced regulation is the US Health Information Portability and Accountability Act (HIPAA) (HIPAA, 1996). According to the HIPAA "Safe Harbor" approach, the clinical records are considered de-identified when 18 categories of PHI are removed, and the remaining information cannot be used alone or in combination with other information to identify an individual. These PHI categories include names, geographic locations, elements of dates (except year), telephone and fax numbers, medical record numbers or any other unique identifying numbers, among others. Since manual removal of PHI by medical professionals proved to be prohibitively time-consuming, tedious, costly and unreliable (Douglass et al., 2004; Neamatullah et al., 2008; Deleger et al., 2013), extracting PHI requires more reliable, faster and cheaper automatic de-identification systems based on Natural Language Processing (NLP) methods (Meystre et al., 2010).

The extraction of PHI can be viewed as a Named Entity Recognition (NER) problem applied in medical domain for the de-identification (Nadeau, 2007). However, even though both traditional NER and the de-identification involve the automatic recognition of particular phrases in text (persons, organizations, locations, dates, etc.), the de-identification differs in important ways from traditional NER (Wellner et al., 2007). In contrast to general NER focused on newspaper texts, the de-identification deals with the clinical narratives characterized by fragmented and incomplete utterances, the lack of punctuation marks and formatting, many spelling and grammatical errors, as well as domain specific terminology and abbreviations. Since the de-identification is the first step towards identification and extraction of other relevant clinical information, it is extremely important to overcome the problem of significantly large number of eponyms and other non-PHI erroneously categorized as PHI. For instance, the anatomic locations, devices, diseases and procedures could be erroneously recognized as PHI and removed (e.g. *"The Zvezdara method"*[1] vs. *Clinical Center "Zvezdara"*), reducing the usability and the overall meaning of clinical notes, and thus the accuracy of subsequent automatic processes performed on the de-identified documents.

In this paper we introduce our automatic clinical narrative text de-identification system, based on a rapid

[1] The original surgical 2-step arteriovenous loop graft procedure developed in Clinical Center "Zvezdara", Belgrade, Serbia. *Zvezdara* is a municipality of Belgrade.

adaptation of the existing NER system for Serbian. The aim of this study is to evaluate the accuracy of PHI removal and replacement while preserving all the medically relevant information about the patient and keeping the resulting de-identified document usable for subsequent information extraction processes.

2. Related work

Over the past twenty years, various text de-identification approaches have been developed, but relatively few published reports are focused only on the unstructured medical data. The extensive review of recent research in the automatic de-identification of narrative medical texts is given in (Meystre et al., 2010). However, most of them are highly specialized for specific document types or a subset of identifiers. Regarding the general nature of applied de-identification methods, the majority of the systems used only one or two specific clinical document types (pathology reports, discharge summaries or nursing progress notes) for the evaluation (Gardner, 2008; Neamatullah et al., 2008; Uzuner et al., 2008; Gardner et al., 2010), while only a few of them were evaluated on a larger scale, with a more heterogeneous document corpus (Sweeney, 1996; Taira, 2002; Ruch at al., 2000; Ferrández et al., 2013). The selection of targeted PHI varied from patient names only (Taira, 2002) to all 17 textual HIPAA PHI categories (Aramaki et al., 2006; Neamatullah et al., 2008; Wellner et al., 2007), or even everything but valid medical concepts (Berman, 2003; Morrison et al., 2009).

The de-identification approaches applied in medical domain are mostly classified into the rule-based or machine learning methods, while some hybrid approaches (Ferrández et al., 2013) efficiently take advantage of both previous methods. The rule-based methods (Neamatullah et al., 2008; Morrison et al., 2009) make the use of dictionaries and hand-crafted rules to identify mentions of PHI, with no annotated training data. Although these systems are often characterized with the limited generalizability that depends on the quality of the patterns and rules, they can be easily and quickly modified by adding rules, dictionary terms or regular expressions in order to improve the overall performance (Meystre et al., 2014). On the other hand, the machine-learning methods (Aramaki et al, 2006; Wellner et al., 2007; Gardner, 2008; Uzuner et al., 2008; Aberdeen et al., 2010), proved to be more easily generalized, automatically learn from training examples to detect and predict PHI. However, these methods require large amounts of annotated data and the adaptation of the system might be difficult due to often unpredictable effects of a change. In 2006, within the Informatics for Integrating Biology and the Bedside (i2b2) project and organized de-identification challenge, a small annotated corpus of hospital discharge summaries were shared among interested participants, providing the basis for the system development and evaluation. Detailed overview and evaluation of the state-of-the-art systems that participated in the i2b2 de-identification challenge is given in (Uzuner et al., 2007).

Aside from systems specifically designed for the de-identification purpose, some NER tools trained on newspaper texts also obtained respectable performance with certain PHI categories (Benton et al. 2011, Wellner et al., 2007).

3. Materials and methods

This section provides an overview of our rule-based de-identification approach for narrative medical texts.

3.1. Training and text corpus

The training corpus for our system development consisted of 200 randomly selected documents from different specialties, generated at three Serbian medical centers. They included discharge summaries (50), clinical notes (50) and medical expertise (100), with a total word count of 143,378. The discharge summaries and clinical notes are unstructured free text typed by the physicians at the conclusion of a hospital stay or series of treatments, including observations about the patient's medical history, his/her current physical state, the therapy administered, laboratory test results, the diagnostic findings, recommendations on discharge and other information about the patient state. Medical expertise documents were oversampled because of their richness in the PHI items.

The characteristic of medical narratives confirmed in our corpus are fragmented and incomplete utterances and lack of punctuation marks and formatting. Moreover, as these documents are usually written in a great hurry there is also an unusual number of spelling, orthographic and typographic errors, much larger than in, for instance, newspaper texts from the Web. For the moment, we have taken these documents as they are and we are not attempting to correct them. In some particular situations we are able to guess the intended meaning, as will be explained in the next section.

3.2. The NER system

The primary resources for natural language processing of Serbian are consisting of lexical resources and local grammars developed using the finite-state methodology as described in (Courtois and Silberztein, 1990; Gross, 1989). For development and application of these resources the Unitex corpus processing system is used (Paumier, 2011). Among general resources used for NER task are the morphological e-dictionaries, covering both general lexica and proper names, as well as simple words and compounds, including not only entries collected from traditional sources, but also entries extracted from processed texts (Krstev et al., 2013). Besides e-dictionaries, for the recognition and morphosyntactic tagging of open classes of simple words and compounds generally not found in dictionaries, the dictionary graphs in the form of finite-state transducers (FSTs) are used. Due to the high level of complexity and ambiguity of named entities, the additional resources for NER were developed. The Serbian NER system is organized as a cascade of FSTs – CasSys (Maurel et al., 2011), integrated in the Unitex corpus processor. Each FST in a cascade modifies a piece of text by replacing it with a lexical tag that can be used in subsequent FSTs. For instance, in a sequence *Dom zdravlja "Milutin Ivković"* 'Health Center 'Milutin Ivković" first a full name 'Milutin Ivković' is recognized and tagged {Milutin Ivković,.NE+persName+full:s1v}, and then a subsequent transducer in the cascade uses this information to appropriately recognize and tag the full organization name (that can also be subsequently used):

(1) {Dom zdravlja "\{Milutin Ivković\, \.NE\+persName \+full\\:s1v\}",.NE+org+:1sq:2sq:7sq:3sq:4sq:5sq:6sq}

Serbian NER system recognizes a full range of traditional named entity types:

- Amount expressions – count, percentage, measurements and currency expressions;
- Time expressions – absolute and relative dates and times of day (fixed and periods), durations and sets of recurring times;
- Personal names – full names, parts of names (first name only, last name only), roles and functions of persons;
- Geopolitical names – names of states, settlements, regions, hydronyms and oronyms;
- Urban names – at this moment only city areas and addresses are recognized.

For the purpose of PHI de-identification not all of these NEs are of interest. For instance, amount expressions should not be de-identified, and roles or functions need not be de-identified. However, we chose not to exclude them from recognition for two reasons: first, if they are recognized correctly that may prevent some false recognition and second, even if they are not of interest for this specific task they may help in recognition of some NEs that are of interest. For instance in Example (2) a name is erroneously typed (both the first and the last name are incorrect) but due to a correct recognition of a person's function the name is also recognized.

(2) *prof. dr sci Bramslav Dimitnjević, specijalista za stomatološku protetiku i ortopediju 'Prof. PhD Bramslav Dimitnjević, a specialist for Prosthetic Dentistry and Orthodontics'*

The finite-state transducers used in the NER cascade use beside general and specific e-dictionaries, as explained before, local grammars that model various triggers and NEs context, such as:

- The use of upper-case letters – for personal names, geopolitical names, organizations, etc.;
- The sentence boundaries – to resolve ambiguous cases where there is not enough other context;
- Trigger words – for instance, *reka* 'river', *grad* 'city' and similar can be used to recognize geopolitical names that are otherwise ambiguous;
- Other type of the context – for instance, a punctuation mark following a country name that coincides with a relational adjective[2] signals that it is more likely a country name than an adjective;
- Other NEs – for instance, an ambiguous city name can be confirmed if it occurs in a list of already recognized NEs representing cities. Also, a five digit number that precedes a name of a city (already recognized) is tagged as a postal code (as used in Serbia).
- Grammatical information – this information is used to impose the obligatory agreement in the case (sometimes also the gender and the number) between the parts of a NE. For instance, in *…istakao je gradonačelnik Londona Boris Džonson…* '…stressed Mayor of London Boris Johnson…' *Londona* can be falsely added to the person's name if grammatical information were not taken into consideration (*Londona* is in the genitive case, while *Boris* and *Džonson* are in the nominative case). This is enabled by grammatical information that is part of NE lexical tags (see Example (1)).

[2] In Serbian many country names coincide with relational adjectives in the feminine gender: *Norveška* 'Norway' and *norveška* 'Norwegian'.

3.3. The PHI de-identification

We used our training corpus for creation and adaptation of patterns that will capture the characteristics of PHI. Through the corpus examination we found that, out of 18 HIPAA PHI categories, only eight appeared in our data. Since there is no annotation standard for PHI tagging, we collapsed some of the HIPAA categories into one (telephone and fax numbers, medical record numbers or any other unique identifying number). In order to maximize patient confidentiality, we adopted a more conservative approach, considering countries and organizations as PHI. For the purposes of this study, we defined the resulting PHI categories as follows:

- Persons (*pers*) – refers to all personal names; includes first, middle and/or last names of patients and their relatives, doctors, judges, witnesses, etc.;
- Dates (*date*) – includes all elements of dates except year and any mention of age information for patients over 89 years of age; according to HIPAA, the age over 89 should be collected under one category 90/120;
- Geographic locations (*top*) – includes countries, cities, parts of cities (like municipalities), postal codes;
- Organizations (*org*) – hospitals and other organizations (like courts);
- Numbers (*num*) – refers to any combination of numbers, letters and special characters representing telephone/fax numbers, medical record numbers, vehicle identifiers and serial numbers, any other unique identifying numbers;
- Addresses (*adrese*) - street addresses.

The processing usually starts with a text having undergone a sentence segmentation, tokenization, part-of-speech tagging and morphological analysis. After general-purpose lexical resources are used to tag text with lemmas, grammatical categories and semantic features, the FST cascade is applied, recognizing persons, functions, organizations, locations, amounts, temporal expressions, etc. Since medical narratives have specific characteristics, the primary issue of date's recognition arose and we added a small cascade of FSTs prior to detection of the sentences. For the de-identification task and the processing of medical data, we performed the adjustments of the temporal expressions FSTs.

The de-identification can be performed in several ways: PHI that needs to be de-identified can be replaced by a tag denoting its corresponding category, with a surrogate text, or both. We have chosen the latter approach. Moreover, since we are dealing with the narrative texts as a result we want to obtain a narrative text as well. To that end, the surrogate text is chosen to agree in the case, gender and number with the PHI it replaces (if applicable). Again, such a replacement is enabled by grammatical information associated with some NE types (personal names, organization names, locations, etc.). In order to preserve the existing interval in days between two events in the text or the duration of specific symptoms, all dates were replaced by a shifted date that is consistent throughout all the de-identified documents.

3.4. An example

In this subsection we will give an example taken from the part of the test corpus containing medical expertise. The

part of one note is given in Example (3).[3] The same expertise after the de-identification and tagging is given in Example (4).[4]

(3) Vaš broj Posl. Br. Ki 250/08
Naš broj 33/06
OPŠTINSKI SUD Istražni sudija G-đa Rada Anđelić-
Vašom naredbom zatražili ste od Komisije lekara veštaka Medicinskog fakulteta Univerziteta u Nišu sudsko medicinsko veštačenje u predmetu Ki 250/08 na okolnost vrste, težine i mehanizma nastanka povreda koje je dana 20.03.2008. god. zadobio oštećeni Marković Ivan iz Dragačeva.
…
PODACI
1. Pri pregledu obavljenom dana 11.08.2007. god. od strane članova Komisije lekara veštaka Medicinskog fakulteta u Nišu Ivana, Mirka, Marković navodi da je rođen 13.05.1986. god. u Dragačevu, živi u Dragačevu, ul. Dositejeva br. 27, po zanimanju elektromehaničar za teničke i rashladne uređaje. Identitet imenovanog utvrđen je na osnovu članovima komisije pokazane lične karte br. 82193. Ivan takođe navodi da je krajem marta meseca 2008. god. oko 1 h posle ponoći sa svojim drugovima sedeo u parku ispred hotela gde je u toku bilo svadbeno veselje.
…
NALAZ
1. U izveštaju doktora Opšte bolnice u Užicu na ime Markoić Ivana, broj protokola 01241, izdatom dana 21.03.2006. god. u 2,30h navedeno je sledeće:" Fractura dens 2 traumatice (dalje nečitko) upućuje se stomatologu radi daljeg lečenja i kvalifikacije povrede "
…
(4) Vaš broj <number PHI="yes">XXXX</number>
Naš <number PHI="yes">XXXX</number>
<org PHI="yes">SUD</org> <pers><role>Istražni sudija gospođa</role> <persName.full PHI="yes">Vilma Kremenko</persName.full></pers>-
Vašom naredbom zatražili ste od
<org PHI="yes">Komisije</org>
<org PHI="yes">fakulteta</org>
<org PHI="yes">Univerziteta</org> sudsko medicinsko veštačenje u predmetu
<number PHI="yes">XXXX</number> na okolnost vrste, težine i mehanizma nastanka povreda koje je dana
<date PHI="yes">26.09.2007.</date> zadobio oštećeni
<persName.full PHI="yes">Barni Kamenko</persName.full> iz
<top.gr PHI="yes">Kamengrada</top.gr>.
…
PODACI
1.{S} Pri pregledu obavljenom dana <date PHI="yes">17.02.2008.</date> od strane članova <org PHI="yes">Komisije</org>
<org PHI="yes">fakulteta</org>
<persName.full PHI="yes">Barni Kamenko</persName.full> navodi da je rođen
<date PHI="yes">19.11.1987.</date> u
<top.gr PHI="yes">Kamengradu</top.gr>, živi u
<top.gr PHI="yes">Kamengradu</top.gr>, <address PHI="yes">ul. Kamenolomska br. 6a</address> , po zanimanju elektromehaničar za teničke i rashladne uređaje.{S} Identitet imenovanog utvrđen je na osnovu članovima komisije pokazane lične karte <number PHI="yes">XXXX</number>.{S} **Ivan** takođe navodi da je krajem <date PHI="yes">septembra 2007.</date> oko 1 h posle ponoći sa svojim drugovima sedeo u parku ispred hotela gde je u toku bilo svadbeno veselje.
NALAZ
<number PHI="yes">XXXX</number>.{S} U izveštaju doktora <org PHI="yes">bolnice</org> na ime <persName.full PHI="yes">**Vilma Kremenko**</persName.full>, broj protokola <number PHI="yes">XXXX</number>, izdatom dana <date PHI="yes">27.09.2007.</date> u 2,30h navedeno je sledeće:{S}" Fractura dens 2 traumatice (dalje nečitko) upućuje se stomatologu radi daljeg lečenja i kvalifikacije povrede "

This example demonstrates our de-identification approach. Some personal data remained: the occurrence of the first name of the patient. Also, the replacement text was not always correct: the male patient's name was once replaced by the female name. These occurrences are bolded and underlined in Example (4).

4. Evaluation results

The previously described system for the automatic de-identification has been evaluated on a set of 100 randomly selected documents (total word count of 35,822), consisting of discharge summaries (60), clinical notes (27) and medical expertise (13). These chosen texts were not used in the system development and present completely unseen material containing many occurrences of PHI. Details about the PHI distribution within the test corpus can be found in Table 1.

PHI/Document type	Cinical reports	Discharge summaries	Medical expertise	Total
pers	52	254	407	713
top	32	219	109	360
org	62	164	242	468
num	20	61	90	171
date	65	133	267	465
adrese	0	64	10	74
Total	231	895	1125	2251

Table 1: The PHI distribution considering document type

The performance has been evaluated with respect to recognition, bracketing and replacement of PHI. For that reason, a new attribute 'check' has been added to each XML tag. Possible values of this attribute were the following:

OK – PHI was correctly recognized, full extent was correctly determined, replacement was correctly assigned;

UOK - UOK1 (PHI type was correctly recognized, but full extent was not correctly determined, some part of PHI was revealed); UOK2 (PHI type was not correctly

[3] This example looks exactly as the original – however, for the purpose of protecting the personal data we have manually replaced all of it.

[4] We wanted to avoid introduction of some real people names and real location names in the de-identified texts. Instead we used names: *Barni Kamenko* (Barney Rubble), *Vilma Kremenko* (Vilma Flintstone), *Kamengrad* (Bedrock), Serbian names for the characters from the sitcom *The Flinstones*, created by Hanna-Barbera Productions, Inc.

determined, but the full extent was correctly determined, PHI successfully masked);

NOK – an utterance tagged falsely as PHI and de-identified;

MISS – PHI was not recognized;

MISS/E – PHI was not recognized because of the incorrect input.

In some cases when it was not so easy to decide which is the most appropriate value for the 'check' attribute (e.g. personal name as a name of an organization), we always treated as correct, for example, personal name tags even though the utterance belongs to organization category.

We report the results of the evaluation using the traditional performance measures: precision (positive predictive value), recall (sensitivity) and F measure (harmonic mean of recall and precision). These measures are calculated at the phrase level, considering the entire PHI annotation as the unit of evaluation.

The harmonic mean of recall and precision is calculated in two ways, using the strict and relaxed criteria. With strict criteria we consider as true positives only fully correctly recognized and de-identified PHI and as false negatives all PHI that were not recognized and de-identified, no matter for what reason (including the incorrect input). With relaxed criteria we consider as true positives all correctly recognized and de-identified PHI including partial recognition and false type attribution, and as false negatives all PHI that were not recognized and de-identified if the input was correct (see Table 2).

	1. Strict criteria	2. Relaxed criteria
TP	OK	OK+UOK
FP	NOK+UOK	NOK
FN	MISS+MISS/E	MISS
P	OK/(OK+NOK+UOK)	(OK+UOK)/(OK+NOK+UOK)
R	OK/(OK+MISS+MISS/E)	(OK+UOK)/(OK+UOK+MISS)

Table 2. Calculation using strict and relaxed criteria: TP (true positive), FP (false positive), FN (false negative), P (Precision), R (Recall)

The overall evaluation of the system is presented in Table 3 and Table 4.

PHI	OK	UOK1	UOK2	MISS	MISS/E	NOK
pers	634	12	47	15	5	30
top	337	0	0	14	9	5
org	434	0	0	28	6	6
num	132	2	0	36	1	8
date	455	4	0	1	5	7
address	63	0	0	1	10	2
Total	2055	18	47	95	36	58

Table 3. Evaluation data

PHI	Precision (p1)	Recall (r1)	F1-measure	Precision (p2)	Recall (r2)	F2-measure
pers	0.88	0.97	0.92	0.96	0.98	0.97
top	0.99	0.94	0.96	0.99	0.96	0.97
org	0.99	0.93	0.96	0.99	0.94	0.96
num	0.93	0.78	0.85	0.94	0.79	0.86
date	0.98	0.99	0.98	0.98	1.00	0.99
Address	0.97	0.85	0.91	0.97	0.98	0.98
Total	**0.94**	**0.94**	**0.94**	**0.97**	**0.96**	**0.97**

Table 4. Performance measures for PHI de-identification

An error analysis shows that every correctly recognized PHI was correctly de-identified. The main source of errors were missed PHI, resulting in the information disclosure. The most missed PHI were numbers and organizations not included in our pattern rules and dictionaries, while fewer than 6% of errors resulted in the revealing of the most sensitive category i.e. person names. Another source of errors that could cause PHI exposure was wrongly determined PHI extent (4.72% of total errors). Fewer than 20% of errors were examples tagged with an incorrect PHI category which may only reduce the readability of the resulting de-identified text without exposing PHI. Since one of the main goals is to preserve medically relevant information, it is important to pay special attention to false positives, which represented 22.83% out of total errors. For *pers*, a majority of false positives were diseases and procedures names.

Our automatic de-identification system achieved very competitive precision and recall rate, showing the overall *F1*-measure of 0.94 (Table 4). High performance was achieved for most PHI types, except for numbers. The highest precision of 0.99 was reached for geographic locations and organizations, followed by dates, addresses and numbers. When partially recognized and wrongly tagged personal names are treated as true positives, the precision of their de-identification is better. With respect to recall, the most important measure for de-identification, dates have the highest rate. Beside dates, almost all PHI categories showed high sensitivity rating from 0.99 to 0.93. The lowest recall rate for numbers (0.78) and addresses (0.85) requires inclusion of missing patterns for these categories. In terms of recall, especially dates and personal names, we may say that our de-identification is sufficient to guarantee high patient privacy, with achieved competitive precision and preserved document usefulness for subsequent applications.

5. Conclusion

In this paper, we described the automatic text de-identification system for medical narrative texts, based on the rapid adaptation of the existing NER system for Serbian. Even though the evaluation of the presented system is conducted on a relatively small set of documents, we have collected the heterogeneous corpus,

consisting of different document types belonging to various medical specialties and institutions. Results of this preliminary evaluation are very promising, indicating that our adapted NER system can achieve high performance on the de-identification task. However, there is still much to be done.

In the future work, we plan to focus on improvements of our strategies, such as completing the existing and adding the new patterns covering the broader formats of PHI (email addresses, URLs, IP address numbers) and the disambiguation of clinical eponyms and abbreviations. Finally, we intent to measure the impact of the de-identification through the subsequent natural language processing task of medical concepts' recognition.

6. References

Aberdeen, J., Bayer, S., Yeniterzi, R., Wellner, B., Clark, C., Hanauer, D., Malin, B. & Hirschman, L. 2010. The MITRE Identification Scrubber Toolkit: Design, training, and assessment. *International Journal of Medical Informatics*, 79**:**849-859.

Aramaki, E., Imai, T., Miyo, K., Ohe, K. Automatic deidentification by using sentence features and label consistency. In: *Workshop on challenges in natural language I2b2 processing for clinical data*. Washington, DC; 2006.

Benton, A., Hill, S., Ungar, L., Chung, A., Leonard, C., Freeman, C. & Holmes, J. H. 2011. A system for de-identifying medical message board text. *BMC Bioinformatics*, 12(Suppl 3):S2.

Berman, J. J. 2003. Concept-match medical data scrubbing - How pathology text can be used in research. *Archives of Pathology & Laboratory Medicine*, 127**:**680-686.

Courtois, B., Silberztein, M. 1990. *Dictionnaires électroniques du français*. Larousse, Paris.

Deleger, L., Molnar, K., Savova, G., Xia, F., Lingren, T., Li, Q., Marsolo, K., Jegga, A., Kaiser, M., Stoutenborough, L. & Solti, I. 2013. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *Journal of the American Medical Informatics Association*, 20:84-94.

Douglass, M., Clifford, G. D., Reisner, A., Moody, G. B., Mark, R. G. 2004. Computer-assisted de-identification of free text in the MIMIC II database. *Computers in Cardiology*, 31:341-344.

Ferrández, O., South, B. R., Shen, S., Friedlin, F. J., Samore, M. H. & Meystre, S. M. 2013. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. *Journal of the American Medical Informatics Association*, 20**:**77-83.

Gardner, J. & Xiong, L. 2008. HIDE: An integrated system for health information DE-identification. In: *Proceedings of the 21st IEEE International Symposium on Computer-Based Medical Systems*. 254-259.

Gardner, J., Xiong, L., Wang, F., Post, A., Saltz, J. & Grandison, T, 2010. An evaluation of feature sets and sampling techniques for de-identification of medical records. In: Veinot T, (ed.), *Proceedings of the 1st ACM International Health Informatics Symposium*. New York:ACM. 183-190.

Gross, M. 1989. The use of finite automata in the lexical representation of natural language. *Lecture Notes in Computer Science*, 377:34-50.

Health Insurance Portability and Accountability Act. P.L. 104-191, 42 USC. 1996.

Krstev, C., Obradović, I., Utvić, M. & Vitas, D. 2014. A system for named entity recognition based on local grammars. *Journal of Logic and Computation*, 24:473-489.

Maurel, D., Friburger, N., Antoine, J. Y., Eshkol-Taravella, I. & Nouvel, D. 2011. Transducer cascades surrounding the recognition of named entities. *Cascades de transducteurs autour de la reconnaissance des entités nommées*, 52**:**69-96.

Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S. & Samore, M. H. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology*, 10:70.

Meystre, S. M., Ferrández, Ó., Friedlin, F. J., South, B. R., Shen, S. & Samore, M. H. 2014. Text de-identification for privacy protection: A study of its impact on clinical text information content. *Journal of Biomedical Informatics*. doi: 10.1016/j.jbi.2014.01.011.

Morrison, F. P., Lai, A. M. & Hripcsak, G. 2009. Repurposing the Clinical Record: Can an Existing Natural Language Processing System De-identify Clinical Notes? *Journal of the American Medical Informatics Association*, 16**:**37-39.

Nadeau, D. & Sekine, S. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30:3-26.

Neamatullah, I., Douglass, M. M., Lehman, L.-W. H., Reisner, A., Villarroel, M., Long, W. J., Szolovits, P., Moody, G. B., Mark, R. G. & Clifford, G. D. 2008. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8:32.

Paumier, S. 2011. Unitex 3.0 User manual. http:// http://www-igm.univ-mlv.fr/~unitex/UnitexManual3.0.pdf.

Ruch, P., Baud, R. H., Rassinoux, A. M., Bouillon, P. & Robert, G. 2000. Medical document anonymization with a semantic lexicon. *Proc AMIA Symp*, 729-733.

Sweeney, L. 1996. Replacing personally-identifying information in medical records, the Scrub system. *Proc AMIA Annu Fall Symp*, 333-337.

Taira, R. K., Bui, A. T. A. & Kangarloo, H. 2002. Identification of patient name references within medical documents using semantic selectional restrictions. *Proc AMIA Annu Symp*, 757-761.

Uzuner, O., Luo, Y. & Szolovits, P. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14**:**550-563.

Uzuner, O., Sibanda, T. C., Luo, Y. & Szovits, P. 2008. A de-identifier for medical discharge summaries. *Artificial Intelligence in Medicine*, 42**:**13-35.

Wellner, B., Huyck, M., Mardis, S., Aberdeen, J., Morgan, A., Peshkin, L., Yeh, A., Hitzeman, J. & Hirschman, L. 2007. Rapidly retargetable approaches to de-identification in medical records. *Journal of the American Medical Informatics Association*, 14**:**564-573.

Procesiranje slovenskega jezika v razvojnem okolju NooJ

Kaja Dobrovoljc

Trojina, zavod za uporabno slovenistiko
Dunajska cesta 116, 1000 Ljubljana
kaja.dobrovoljc@trojina.si

Povzetek

Z razvojem področja gradnje in označevanja obsežnih računalniških zbirk avtentičnih besedil se tudi v slovenskem jezikoslovju povečuje število raziskav, ki temeljijo na njihovem preučevanju. Kljub vse večjemu uveljavljanju korpusnih metod pa trenutno v slovenističnem prostoru obstaja precejšen razkorak med mnogimi raziskovalnimi priložnostmi, ki jih obstoječi korpusi ponujajo, ter njihovo dejansko izrabo. To stanje je v določeni meri tudi posledica pomanjkanja enostavnih orodij za kompleksnejšo obdelavo korpusnih besedil. Kot primer računalniško zmogljivega, a jezikoslovnemu uporabniku prijaznega orodja v prispevku predstavljamo NooJ, jezikoslovno razvojno okolje za izdelavo obsežnih formaliziranih opisov naravnih jezikov in njihovo uporabo v besedilnih korpusih. Na primeru izbranih jezikovnih virov in pravil iz pilotnega modula za slovenščino predstavimo najpomembnejše funkcionalnosti tega razvojnega okolja in prednosti njegovega povezovanja z že obstoječimi viri in orodji za slovenščino.

Slovene Language Processing with NooJ

With continued development in the field of corpus compilation and annotation, there is also an increase in the quantity of linguistic research dealing with their analysis. However, despite the growing recognition of corpus-based methods in Slovene language studies, there exists a considerable gap between many research opportunities arising from the available corpora and their actual realization. To some extent, this discrepancy can be attributed to the lack of user-friendly tools for complex corpus processing. As an example of such computationally powerful and yet linguist-friendly tool, we present NooJ, a linguistic development environment for construction of large-coverage formalized descriptions of natural languages and their application to corpora. By drawing on the examples from the initial module for Slovene, we present some of its most important features, and the prospects of its integration with other existing resources and tools for Slovene language processing.

1. Uvod

Z razvojem področja gradnje in obdelave obsežnih računalniških besedilnih zbirk se povečuje tudi število jezikoslovnih raziskav, ki temeljijo na njihovem preučevanju. Tudi v slovenskem jezikoslovju korpusni pristop postaja vse bolj uveljavljena metoda raziskovanja, toda zdi se, da so raziskovalci pogosto ujeti med načelno zavezo k raziskovanju jezikovne rabe v obsežnih zbirkah avtentičnih besedil na eni strani in nezmožnostjo kompleksne računalniške obdelave, ki jo taka analiza zahteva, na drugi. Posledično se raziskovalci pri načrtovanju metodologije pogosto omejujejo na preučevanje manjših (pod)korpusov, ki omogočajo obvladljivejšo, ročno analizo, ali pa predmet svojega raziskovanja prilagajajo omejenim funkcionalnostim spletnih korpusnih vmesnikov.

Naprednejši prostodostopni spletni konkordančniki, kot sta NoSketchEngine[1] in CUWI[2] (Erjavec, 2013), sicer omogočajo oblikovanje zapletenejših korpusnih poizvedb v obliki regularnih izrazov, a ti pri opisovanju kompleksnejših struktur hitro postanejo težko obvladljivi, prednastavljeni pristopi k analizi izluščenih konkordanc pa se v omenjenih orodjih osredotočajo predvsem na distribucijsko analizo in merjenje besedne povezovalnosti. V slovenskem prostoru tako obstaja realna potreba po orodju, ki bi tudi raziskovalcem brez znanja programiranja omogočilo enostavno opisovanje, luščenje, označevanje in urejanje kompleksnih jezikovnih pojavov na različnih ravneh.

Kot primer takega računalniško zmogljivega, a jezikoslovnemu uporabniku prijaznega orodja v prispevku predstavljamo razvojno okolje NooJ, za katerega je bil pred kratkim razvit tudi pilotni modul slovenski jezik (Dobrovoljc, 2014a).

2. Programska oprema

NooJ je jezikoslovno razvojno okolje za izdelavo obsežnih formaliziranih opisov naravnih jezikov in njihovo uporabo v besedilnih korpusih. Opisi naravnih jezikov so formalizirani v obliki elektronskih slovarjev in na grafu temelječih slovnic (pravil), s katerimi lahko na razmeroma preprost način opisujemo jezikovne pojave na različnih ravneh površinske zgradbe besedil, od besednih oblik do zapletenejših skladenjskih in besedilnih enot.

Za opis jezikovnih prvin in razčlenjevanje besedil NooJ med drugim uporablja končne pretvornike (*finite-state transducers*, FST) za prepoznavanje in označevanje nizov črk in/ali besed, končne avtomate (*finite-state automata*, FSA) za korpusna poizvedovanja, rekurzivne mreže prehodov (*recursive transition networks*, RTN) za izdelavo slovnic z več povezanimi grafi končnih stanj ter napredne rekurzivne mreže prehodov (*enhanced recursive transition networks*, ERTN), ki z uporabo spremenljivk in omejitev omogočajo raznorazne besedilne pretvorbe.

NooJ je bil kot nadgradnja okolja INTEX[3] (Silberztein, 1993) pod avtorstvom istega razvijalca izhodiščno izdelan v ogrodju .NET, v okviru projekta CESAR pa je bila izdelana tudi njegova odprtokodna različica v ogrodju Java (Silberztein, Váradi, Tadić, 2012). Obe različici z grafičnim vmesnikom sta za prenos na voljo na uradni spletni strani orodja.[4] Priročnik (Silberztein, 2003: 206-

[1] http://nl.ijs.si/noske/.
[2] http://nl.ijs.si/cuwi/.

[3] Kot odgovor na izhodiščno zaprtost sistema INTEX je bil leta 2002 izdelan zelo podoben odprtokodni sistem Unitex: http://www-igm.univ-mlv.fr/~unitex/.
[4] http://www.nooj4nlp.net/.

211) opisuje tudi različico za uporabo v ukazni vrstici in programskih vmesnikih (noojapply), ki pa ni javno objavljena, saj se po objavi odprtokodne različice program ne posodablja več.

Na spletni strani so objavljeni tudi jezikovni moduli, tj. zbirke jezikovnih virov (korpusov, leksikonov in slovnic) za procesiranje posameznih jezikov. V nasprotju s programsko opremo je odločitev o dostopnosti in rednem posodabljanju vsebin modulov prepuščena njihovim avtorjem, zato se obseg, kakovost in dostopnost virov za posamezne jezike pogosto precej razlikujejo. Trenutni nabor modulov vključuje 23 jezikov raznolikih oblikoslovnih tipov, pisav in jezikovnih družin, med njimi tudi južnoslovanske (bolgarščina, hrvaščina, srbščina, slovenščina).

V nadaljevanju na izbranih primerih iz obstoječe odprtokodne različice modula za slovenščino[5] predstavimo najpomembnejše značilnosti tega razvojnega okolja, in sicer možnost uvoza označenih in neoznačenih korpusov (3), leksikalne podatkovne zbirke (4), grafični vmesnik za izdelavo oblikoslovnih in skladenjskih pravil (5), možnosti označevanja besedil (6) in konkordančnik za njihovo analizo (7). Te in druge funkcionalnosti orodja NooJ so podrobneje in z več slikovnimi ponazoritvami predstavljene v uradnem priročniku za uporabnike (Silberztein, 2003), primeri konkretne uporabe orodja za različne namene obravnave naravnega jezika in nekatere novejše funkcionalnosti, s katerimi bi veljalo posodobiti obstoječi priročnik, pa so predstavljeni v zbornikih vsakoletnih konferenc (NooJ International Conference).

3. Korpusi

NooJ lahko procesira eno (.not) ali več besedilnih datotek (.noc) v več kot 150 različnih vhodnih formatih (npr. html, MS-Word, pdf, rtf itd.) in različnih načinih kodiranja. Uporabniki lahko besedila ustvarijo v integriranem urejevalniku ali pa jih vanj uvozijo kot zunanje datoteke, pri čemer je treba v obeh primerih določiti tudi izbrani razmejevalnik besedil (npr. odstavek, XML element ali določen izraz v datoteki).

NooJ poleg golih, neoznačenih besedil omogoča tudi uvoz že označenih besedil (v formatu XML), kar pomeni, da lahko uporabniki skupaj z besedilom v program uvozijo tudi podatke o njegovi strukturi, prevodu, slovničnih oznakah ipd. Pri uvozu datoteke v formatu XML v NooJ se imena elementov znotraj izbranega razmejevalnika besedila avtomatsko pretvorijo v skladenjske oz. semantične oznake (ime elementa postane ime skladenjske oz. semantične kategorije za vsebino elementa). Kot bomo videli v nadaljevanju, je izjema zgolj posebni element <LU>, ki napoveduje osnovne jezikovne enote (besedne pojavnice).

Z namenom preizkusa in prikaza uvoza besedil z različnimi tipi metapodatkov smo v slovenski modul vključili tri obstoječe korpuse (tabela 1): neoznačeno besedilo romana Telesni čuvaj (Mazzini, 2000), korpus ccKres (Logar Berginc et al., 2012), označen s statističnim označevalnikom Obeliks (Grčar et al., 2012), in ročno označeni korpus ssj500k (ibid.).

Korpus	Št. bes. pojavnic	Označenost
Telesni čuvaj.not	78.367	neoznačen
ssj500k.not	500.295	oblike, skladnja, NER
ccKres.noc	10.000.532	oblike

Tabela 1: Velikost in označenost korpusov v slovenskem modulu za NooJ.

Korpusa ssj500k in ccKres sta v svoji izhodiščni različici zapisana v formatu XML TEI P5[6], kakršnega predvidevajo krovne specifikacije za zapis korpusov, razvitih v okviru projekta Sporazumevanje v slovenskem jeziku[7]. Kot prikazuje slika 1, smo morali pred uvozom v NooJ izhodiščni format omenjenih korpusov prilagoditi tako, da smo besedne pojavnice (v procesu tokenizacije označene z <w>) pretvorili v format, na podlagi katerega NooJ vsebino elementa <LU> prepozna kot besedno obliko, vrednost neobveznega atributa LEMMA kot lemo in vrednost obveznega atributa CAT kot slovnično kategorijo (besedno vrsto). Morebitni drugi metapodatki o pojavnici lahko tema dvema sledijo v obliki poljubno poimenovanih atributov in njihovih vrednosti; v našem primeru je to informacija o celotni oblikoskladenjski oznaki (atribut MSD), pri skladenjsko razčlenjenem korpusu pa tudi o identifikatorju pojavnice v stavku (atribut ID), skladenjski odnosnici (HEAD) in skladenjskem razmerju (DEPREL).

```
<s id="ssj1.1.5">
   <LU LEMMA="ta" CAT="P" MSD="Pd-nsg" ID="t1" HEAD="t5" DEPREL="Obj">Tega</LU>
   <LU LEMMA="se" CAT="P" MSD="Px------y" ID="t2" HEAD="t5" DEPREL="PPart">se</LU>
   <LU LEMMA="sploh" CAT="Q" MSD="Q" ID="t3" HEAD="t0" DEPREL="Root">sploh</LU>
   <LU LEMMA="biti" CAT="V" MSD="Va-r1s-y" ID="t4" HEAD="t5" DEPREL="PPart">nisem</LU>
   <LU LEMMA="zavesti" CAT="V" MSD="Vmep-sm" ID="t5" HEAD="t0" DEPREL="Root">zavedel</LU>
   <c ID="t6" HEAD="t0" DEPREL="Root">.</c>
</s>
```

Slika 1: Stavek korpusa ssj500k v formatu NooJ XML.

4. Slovarji

Leksikoni oz. slovarji (dictionaries, .nod) v orodju NooJ opravljajo vlogo podatkovnih zbirk za prepoznavanje eno- in večbesedne leksike ter opisovanje njihovih oblikoslovnih, skladenjskih, pomenskih in drugih lastnosti. Tipična leksikonska iztočnica je opisana kot niz leme, besednovrstne oznake ter drugih lastnosti (slika 2). Njihov nabor in poimenovanja poljubno določajo uporabniki, v slovarju pa so lahko leksikalni enoti pripisane kot značilke (npr. +splošni) ali kot niz lastnosti in njene vrednosti (+vrsta=splošni). Vrednosti atributov lahko vsebujejo metajezikovne informacije (npr. +vrsta=splošni), besedilo v izhodiščnem jeziku (npr. +sinonim=zlahka oz. +sinonim="brez napora") ali besedilo v tujem jeziku (npr. +ANG=effortlessly).

```
lahko,R+Type=general+FLX=LAHKO
nizko,R+Type=general+FLX=LAHKO
ozko,R+Type=general+FLX=LAHKO
```

Slika 2: Tipični zapis iztočnic v leksikonih NooJ na primeru prislovov iz leksikona Sloleks.

[5] http://www.nooj4nlp.net/pages/slovene.html.

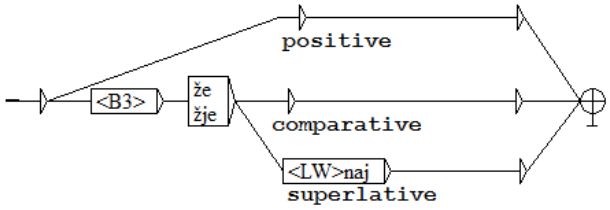
[6] www.tei-c.org/P5/Guidelines/.

[7] http://www.slovenscina.eu/.

4.1. Vzorci pregibanja

Med leksikalnimi atributi s posebnim privzetim pomenom sta najpomembnejša +FLX in +DRV, ki leksikonsko iztočnico (lemo) povezujeta s pripadajočim pregibnim oz. besedotvornim pravilom. Pravila za pregibanje so formalizirana v ločenih datotekah (.nof), in sicer bodisi v obliki besedila bodisi v obliki grafa. Vrednost atributa +FLX (vedno zapisana z velikimi črkami) tako iztočnico v leksikonu poveže z enako poimenovanim pravilom, ki v procesu transformacije osnovne oblike z dodajanjem črk in/ali uporabo posebnih operatorjev (npr. za brisanje, premikanje naprej/nazaj, podvajanje znakov ipd.) iz osnovne ustvari ustrezne pregibne oblike in jim pripiše morebitne dodatne lastnosti.

Slika 3 tako prikazuje opis pregibnega vzorca LAHKO za prislove z variantnim stopnjevanjem (npr. *lahko-laže/lažje-najlaže/najlažje*). Vzorec določa, da je oblika za nedoločno stopnjo enaka lemi, obliki za primerniško in presežniško obliko pa se tvorita tako, da se osnovni obliki brez zadnjih treh črk (<B3>) dodata končnici *-že* in *-žje*, pri čemer se na začetek obeh presežniških variant doda še predpona *naj-* (<LW>naj).




```
LAHKO = <E>/=positive | <B3>že/=comparative |
<B3>žje/=comparative | <B3>že<LW>naj/=superlative
| <B3>žje<LW>naj/=superlative;
```

Slika 3: Primer pregibnega vzorca LAHKO za obrazilno stopnjevanje prislovov v grafični (zgoraj) in besedilni obliki (spodaj).

Pri poskusu prenosa referenčnega oblikoslovnega leksikona Sloleks[8], ki v formatu XML LMF opisuje več kot 100.000 lem s pripadajočimi pregibnimi oblikami, se je kmalu izkazalo, da izluščeni oblikoslovni vzorci vsebujejo precej nedoslednosti, napak in neskladnosti z jezikovno rabo, zato je pred njihovo formalizacijo in povezovanjem z iztočnicami (lemami) nujna še dodatna faza ročne evalvacije. Predlog cevovodnega procesa za polavtomatsko validacijo oblikoslovnih vzorcev za slovenščino smo že preizkusili na primeru obrazilnega pregibanja prislovov (Dobrovoljc, 2014b), rezultati pa so v obliki formaliziranih vzorcev ter posodobljenega in razširjenega leksikona prislovov vključeni tudi v trenutni modul.

Do zaključka procesa sistematične evalvacije vzorcev drugih besednih vrst je celoten leksikon Sloleks v trenutno različico modula zato vključen v obliki, ki ne predvideva povezave s konkretnim formaliziranim vzorcem, temveč posamične oblike z vsemi oblikoskladenjskimi lastnostmi[9] navaja kot ločene leksikonske enote (slika 4). To vpliva na hitrost, ne pa tudi na rezultat procesiranja, saj pripisane oznake ne glede na format vsebujejo enak nabor informacij o slovničnih lastnosti oblik.

```
garažisti,garažist,N+Type=common+Gender=m
asculine+Number=plural+Case=instrumental
```

Slika 4: Primer zapisa iztočnice leksikona Sloleks, pretvorjenega v format za uporabo v orodju NooJ.

5. Slovnice

Drugi temeljni vir jezikovnih opisov v orodju NooJ so slovnice (grammars), ki so v nasprotju z opisovanjem končnega nabora leksike v slovarjih namenjene opisovanju pravil za produktivnejše slovnične pojave na vseh jezikovnih ravneh, pri tem pa se na različne načine povezujejo tudi z leksikalnimi informacijami. Poleg že omenjenih slovnic za pregibanje in besedotvorje, ki opisujejo pregibne lastnosti leksikonskih iztočnic, NooJ vključuje tudi vmesnika za oblikovanje oblikoslovnih pravil (morphological grammars), ki opisujejo morfemske lastnosti besednih oblik, in skladenjskih pravil (syntactic grammars), ki opisujejo skladenjske in semantične lastnosti eno- ali večbesednih izrazov.

Uporabniki oba tipa pravil opisujejo v obliki eno- ali večnivojskih grafov, pri katerih niz vozlišč med začetnim in končnim vozliščem grafa označuje bodisi zaporedje črk ali morfemov znotraj pojavnice (oblikoslovne slovnice) bodisi zaporedje ene ali več pojavnic (skladenjske slovnice). Oba tipa pravil v svojih vozliščih in njihovih oznakah omogočata uporabo spremenljivk in nekaterih privzetih operatorjev, npr. za male in velike začetnice, naglašene in nenaglašene črke, dolžino besede, začetek ali konec stavka ipd.

Vmesnik za izdelavo slovničnih grafov vsebuje tudi nekaj koristnih funkcij, ki uporabnikom na vizualno razumljiv način omogočajo sprotno validacijo pravil, npr. preverjanje seznama pozitivnih in negativnih primerov, generiranje vseh možnih poti grafa (oblik, skladenjskih vzorcev ipd.) in razhroščevanje.

5.1.1. Oblikoslovne slovnice

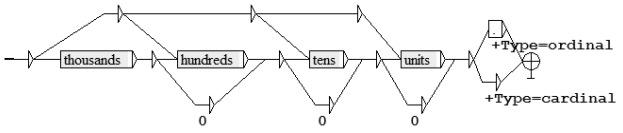
Kot smo že omenili v poglavju o leksikalnih zbirkah, se besedne oblike privzeto analizirajo z vpogledom v leksikon, ki obliko povezuje s pripadajočimi slovničnimi in drugimi lastnostmi. Vendarle pa je nekatere besedne oblike namesto eksplicitnega zapisovanja v leksikalnih zbirkah smiselneje opisovati s posebnimi generičnimi pravili, ki hkrati opisujejo več različnih oblik z enakimi slovničnimi lastnostmi, kot so npr. števniki. Oblikoslovne slovnice (.nom) so tako namenjene predvsem opisovanju tovrstnih produktivnih oblikotvornih pravil za besedne oblike s podobnimi lastnostmi. Najpogosteje se uporabljajo za poenotenje oblikovnih variant, prepoznavanje in označevanje neologizmov, ustvarjanje povezav med besedotvorno povezanimi besedami ipd.

Med oblikoslovnimi slovnicami v slovenskem modulu v prispevku izpostavljamo dve, ki prikazujeta dva možna namena njihove uporabe. Prva slovnica (slika 5) prepoznava rimske številke, določa njihove slovnične

[8] http://www.slovenscina.eu/sloleks/opis.

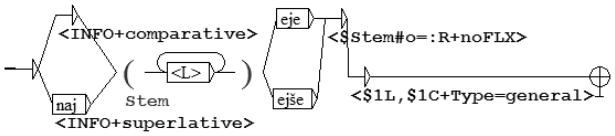
[9] Oblikoslovni metapodatki vseh omenjenih virov temeljijo na naboru oznak, razvitem v okviru projektov MULTEXT-East (Erjavec, 2010) in JOS (Erjavec in Krek, 2008). S tega vidika so slovenski viri kompatibilni tudi z viri drugih jezikov, vključenih v projekt MULTEXT-East, ki imajo obenem razvite tudi svoje module znotraj okolja NooJ (Stanković et al., 2012).

lastnosti (zapis, vrsta) in jih pretvarja v ekvivalentni arabski zapis (npr. *MMLXIV* v *2064*). To je tipični primer pravila, s katerim na hiter in preprost način opišemo večje skupine besednih oblik (npr. vse možne rimske številke), med katerimi so v leksikone običajno vključene le najpogostejše.



Slika 5: Oblikoslovna slovnica za prepoznavanje in pretvarjanje rimskih številnikov.

Druga oblikoslovna slovnica ponazarja primer pravila za procesiranje neznanih besednih oblik, ki smo ga uporabili pri evalvaciji leksikona Sloleks. Slovnica namreč preverja, ali se za prislove, ki so v izhodiščnem leksikonu označeni kot nestopnjevani (npr. *zavzeto*), v rabi pojavljajo tudi obrazilno stopnjevane (primerniške in presežniške) oblike (npr. *zavzeteje, najzavzeteje*). Slovnični graf (slika 6) tako poišče neznane besedne oblike, ki se končajo z enim od obeh možnih obrazil (*-eje* ali *-ejše*), in preveri, ali je izluščeni koren (spremenljivka $Stem) z dodano končnico -o ($Stem#o) v izhodiščnem slovarju označen kot nestopnjevani prislov (:R+noFLX). Če je pogoj izpolnjen, besedna oblika podeduje lemo ($1L) in besedno vrsto ($1C) prislova v leksikonu, doda pa se ji oznaka za primernik oz. presežnik.



Slika 6: Slovnica za prepoznavanje obrazilno stopnjevanih prislovov.

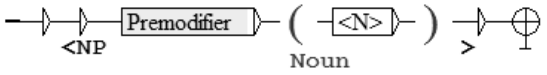
5.1.2. Skladenjske slovnice

Skladenjske ali lokalne slovnice (.nog) so namenjene opisovanju in označevanju eno- ali večbesednih izrazov. Tipično se uporabljajo za skladenjsko in pomensko razčlenjevanje, luščenje in označevanje imenskih entitet ter drugih prekinjenih ali neprekinjenih stalnih besednih zvez, pa tudi za avtomatsko razdvoumljanje besednih oblik v kontekstu ter besedilne pretvorbe, kot sta npr. parafraziranje in prevajanje.

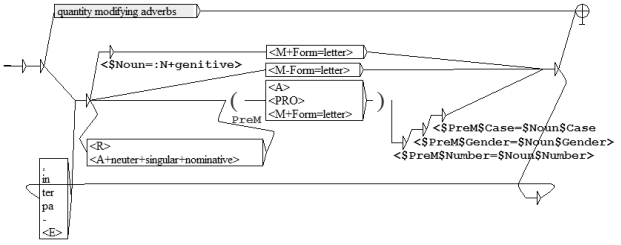
V primerjavi z drugimi na pravilih temelječimi aplikacijami za računalniško obdelavo besedil je ena izmed glavnih prednosti okolja NooJ dejstvo, da sta skladenjska in oblikoslovna raven medsebojno povezljivi. To pomeni, da lahko uporabniki v skladenjska pravila z različnimi operatorji za spremenljivke vključujejo tudi raznorazne pretvorbe njihovih vrednosti (npr. lematizacijo, pregibanje, priklic določene leksikalne lastnosti), določajo omejitve (npr. pogoj ujemanja v izbranih slovničnih lastnostih) in spreminjajo njihovo zaporedje.

Slovenski modul vsebuje nekaj osnovnih slovnic za prepoznavanje in označevanje različnih vrst imenskih entitet (npr. lastnih imen, besednih številnikov, datumov in časovnih izrazov) ter nekaterih površinskih skladenjskih struktur. Primer druge skupine je denimo slovnica, ki v oblikoskladenjsko označenih besedilih prepoznava podredno zložene samostalniške besedne zveze, tj. zveze samostalnika in njegovih (nestavčnih) določil oz. prilastkov.

Na najvišji ravni pravila (slika 8) je besedna zveza opredeljena kot samostalnik, pred katerim stoji eno ali več določil. Vdelani podgraf (slika 9) pa nato podrobneje določa možne strukture določil, tj. končen nabor količinskih prislovov (*nekaj, malo, veliko* ipd.) ter neomejen niz številnikov in/ali prilastkov (spremenljivka $PreM), ki se morajo z jedrnim samostalnikom ujemati v spolu, sklonu in številu. Ujemalni prilastki imajo lahko tudi sami določila v obliki prislovov ali pridevnikov v imenovalniku srednjega spola ednine (npr. *slovensko* v zvezi *slovensko-francoski odnosi*).



Slika 7: Slovnica za prepoznavanje podredno zloženih samostalniških besednih zvez (prvi nivo).



Slika 8: Slovnica za opis levih prilastkov v podredno zloženih samostalniških besednih zvezah (drugi nivo).

6. Označevanje

V prispevku smo že večkrat nakazali, da oblikoslovne in skladenjske slovnice niso namenjene zgolj prepoznavanju jezikovnih pojavov, temveč tudi njihovemu označevanju.

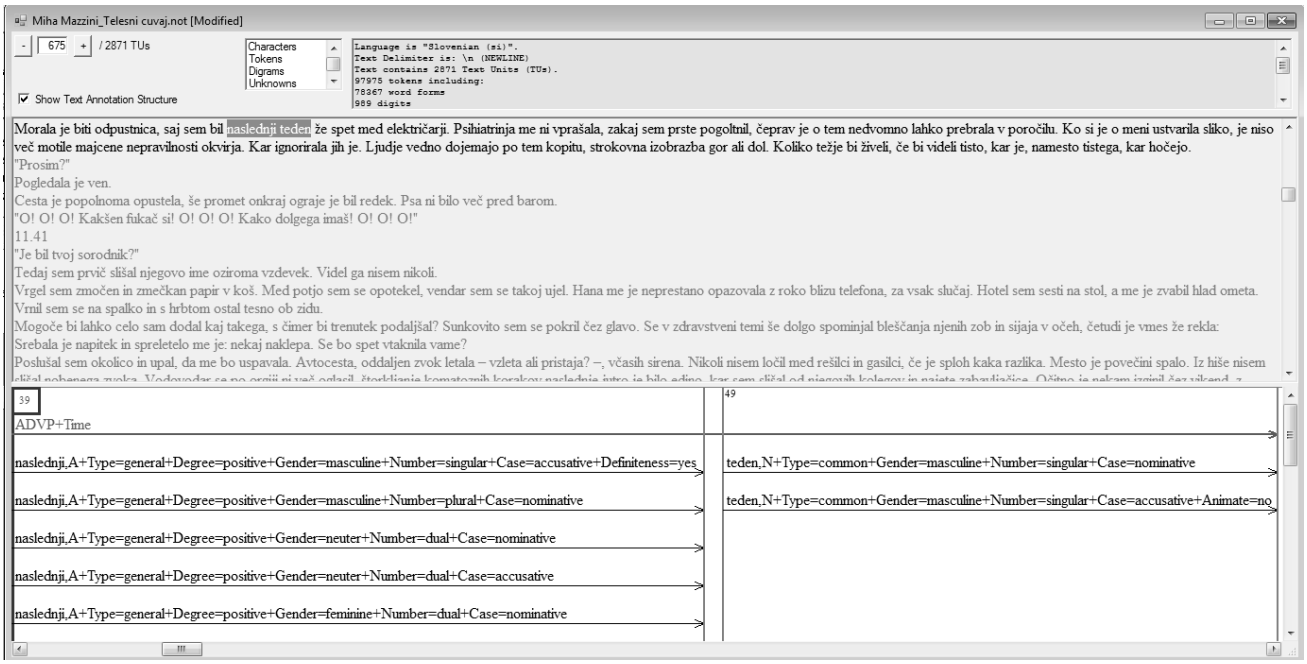
Pri označevanju besedil NooJ ustvarja pare (*pozicija, informacija*), ki označujejo, da ima določen niz v besedilu določene lastnosti. Te binarne oznake se v sistemu shranijo v t. i. strukturo označenega besedila (text annotation structure, TAS), pri čemer se ta nenehno usklajuje z izhodiščno besedilno datoteko, ki se sama nikoli ne spreminja. Binarne oznake se lahko nanašajo tako na posamezne besede (npr. označevanje besede *miza* s kategorijo samostalnika), na del besede (npr. razčlenjevanje pojavnice *karkoli*) ter na neprekinjene in prekinjene večbesedne enote (npr. *okrogla miza* ali *potegnti* nekaj *iz klobuka*). Poleg dodajanja oznak v strukturo označenega besedila NooJ omogoča tudi izvoz označenih besedil in uvoz že označenih besedil (v datotekah XML).

NooJ pri označevanju vedno vrne vse možne oznake, a obenem omogoča tudi njihovo nadaljnje odstranjevanje (razdvoumljanje) na avtomatski (s pravili v obliki

skladenjskih slovnic), polavtomatski (s filtriranjem po seznamu konkordanc) ali ročni način.

Uporabnik pred kakršnimkoli označevanjem v nastavitvah sam izbire relevantne vire (slovarje in slovnice) za jezikoslovno analizo. Določa lahko tudi njihovo zaporedje (stopnjo pomembnosti), pri čemer se nižje uvrščeni viri upoštevajo zgolj pri analizi pojavov, ki jih višje uvrščeni viri niso obravnavali. Ta mehanizem se tako tipično uporablja predvsem za procesiranje neznanih besed (z viri, uvrščenimi za leksikone) oz. za popravke v tokenizaciji večbesednih enot (z viri, uvrščenimi pred leksikone).

Vse informacije v strukturi označenega besedila so v korpusnem vmesniku vizualizirane pod besedilom (slika 9), pri čemer so leksikalne in skladenjske oznake barvno ločene, uporabnik pa jih lahko v primeru ročnega razdvoumljanja tudi ureja.

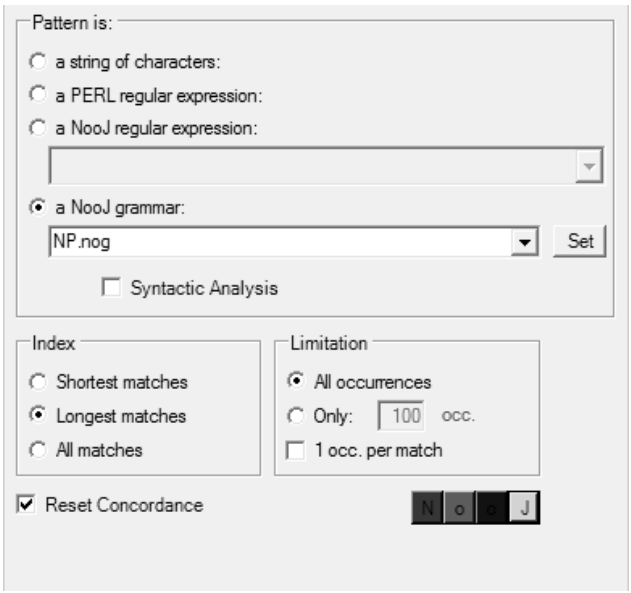


Slika 9: Prikaz strukture označenega besedila z nerazdvoumljenimi oblikoskladenjskimi oznakami po aplikaciji skladenjske slovnice za prepoznavanje časovnih izrazov.

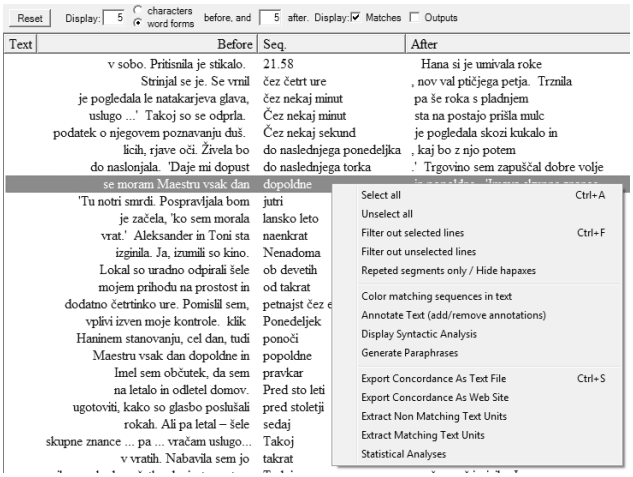
7. Konkordančnik

Poleg zmogljivega opisovanja in procesiranja besedil NooJ odlikuje tudi vmesnik za luščenje korpusnih konkordanc. Kot prikazuje slika 10, lahko uporabniki po besedilu in njegovih raznolikih oznakah poizvedujejo s črkovnimi nizi, različnimi tipi regularnih izrazov in tudi z neposrednim vnosom (skladenjskih) slovnic, ki se tako uporabljajo tudi oz. predvsem za izdelavo kompleksnih korpusnih poizvedb, ki ne spreminjajo označevalne strukture.

Uporabniki lahko priklicane konkordance (slika 11) poljubno urejajo, filtrirajo in izvažajo, jim dodajajo ali odstranjujejo oznake (npr. pri polavtomatskem razdvoumljanju), ogledajo pa si lahko tudi nekatere statistične izračune (pogostost, standardna vrednost, relevantnost za posamezno besedilno enoto, podobnost besedišča).



Slika 10: Vmesnik za iskanje po korpusu, v katerem je kot iskalni pogoj vneseno pravilo za prepoznavanje samostalniških besednih zvez.



Slika 11: Konkordančni niz za skladenjsko slovnico, ki prepoznava časovne izraze.

8. Zaključek

V prispevku smo na omejenem naboru primerov jezikovnih virov in pravil iz pilotnega modula za slovenščino predstavili nekaj temeljnih značilnosti razvojnega okolja NooJ. V primerjavi s splošnimi konkordančniki in drugimi vmesniki za analizo korpusnih besedil NooJ jezikoslovnim in drugim raziskovalcem ponuja možnost naprednejše obdelave korpusnih besedil, ne da bi ti za to potrebovali napredno računalniško predznanje. Odlikujeta ga predvsem vmesnik za razmeroma preprost opis raznolikih jezikovnih pojavov v obliki grafov ter možnost njihove takojšnje uporabe na korpusnih besedilih.

Čeprav je NooJ prvenstveno namenjen razvoju samostojnih, na pravilih temelječih orodij za strojno označevanje jezika, menimo, da se znotraj slovenskega prostora njegov največji potencial skriva v povezovanju z drugimi, že obstoječimi jezikovnimi viri in orodji za strojno procesiranje slovenščine.

V prvi vrsti imamo v mislih možnost oblikovanja kompleksnih korpusnih poizvedb po površinski in označeni strukturi besedila, denimo za luščenje podatkov iz površinskoskladenjsko razčlenjenih korpusov, govornih korpusov ali drugih korpusov, ki poleg slovničnih lastnosti besednih oblik vsebujejo tudi druge vrste in ravni jezikoslovnih oznak.

Druga obetavna možnost uporabe orodja NooJ je v približevanju obstoječih korpusnih virov jezikoslovnim raziskovalcem, ki so te doslej zaradi označevalnih napak ali nestrinjanja z označevalnim sistemom pogosto zavračali kot nezanesljive. Funkcionalnosti orodja NooJ omogočajo preprosto izdelavo hibridnih orodij za nadgradnjo ali dopolnitev oznak v izhodiščnih virih, npr. s pravili za usmerjeno odpravljanje napak, prilagajanje specifičnim raziskovalnim potrebam oz. teoretskim nazorom ter druge oblike hevrističnega usmerjanja statističnih jezikovnih modelov.

Nenazadnje NooJ kot odprtokodna programska oprema predstavlja tudi priročno komunikacijsko stičišče jezikoslovne in računalniške skupnosti, saj jezikoslovcem omogoča preprosto formalizacijo opazovanih jezikovnih pojavov, informatikom pa njihovo brezšivno implementacijo v širše računalniške sisteme.

Literatura

Dobrovoljc, K., 2014a. Introduction to Slovene Language Resources for NooJ. V S. Koeva, S. Mesfar in M. Silberztein (ur.), *Formalising Natural Languages with NooJ 2013: Selected Papers from the NooJ 2013 International Conference*. Newcastle: Cambridge Scholars Publishing. 27-40.

Dobrovoljc, K., 2014b. Re-evaluating morphological dictionaries: the case of adverbs in Slovene. *NooJ 2014 International Conference*. [v objavi]

Erjavec, T. in S. Krek, 2008. Oblikoskladenjske specifikacije in označeni korpusi JOS. V: T. Erjavec in J. Žganec Gros (ur.): *Zbornik Šeste konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 49-53.

Erjavec, T., 2010. MULTEXT-East version 4: multilingual morphosyntactic specifications, lexicons and corpora. V: N. Calzolari (ur.): *Proceedings of the 7th International Conference on Language Resources and Evaluations, 19-21 May 2010, Valletta, Malta*. 2544–2547.

Erjavec, T., 2013. Korpusi in konkordančniki na strežniku nl.ijs.si. *Slovenščina 2.0*, 1:24-49.

Grčar, M, S. Krek in K. Dobrovoljc, 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V T. Erjavec in J. Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 89-94.

Logar Berginc, N., M. Grčar, M. Brakus, T. Erjavec, Š. Arhar Holdt in S. Krek, 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba.* Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Mazzini, M., 2000. *Telesni čuvaj: verzija 1.72.* Ljubljana: Študentska založba.

Silberztein, M., 1993. *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*. Pariz: Elsevier Masson.

Silberztein, M., 2003. *NooJ Manual*. Dostopno na: http://www.nooj4nlp.net/NooJManual.pdf.

Silberztein, M., T. Váradi in M. Tadić, 2012. Open source multi-platform NooJ for NLP. *Proceedings of COLING 2012: Demonstration Papers.* 401-408.

Stanković, R., M. Utvić, D. Vitas, C. Krstev in I. Obradović, 2012. On the Compatibility of Lexical Resources for NooJ. V: K. Vučković, B. Bekavac, & M. Silberztein (ur.): *Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the 2011 International Nooj Conference*. Cambridge Scholars Publishing. 96–109.

Named Entity Recognition in Croatian Tweets

Krešimir Baksa, Dino Dolović, Goran Glavaš, Jan Šnajder

University of Zagreb, Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
{kresimir.baksa, dino.dolovic, goran.glavas, jan.snajder}@fer.hr

Abstract

Existing named entity extraction tools, typically designed for formal texts written in standard language (e.g., news stories, essays, or legal texts), do not perform well on user-generated content (e.g., tweets). In this paper we present a supervised approach for named entity recognition and classification for Croatian tweets. Comparison of three different sequence labeling models (HMM, CRF, and SVM) revealed that CRF is the best model for the task, achieving a micro-averaged F_1-score of over 87%. We also demonstrate that the state-of-the-art NER model designed for Croatian standard language texts performs much worse than our Twitter-specific NER models.

Prepoznavanje imenskih entitet v hrvaških tvitih

Obstoječa orodja za prepoznavanje imenskih entitet, ki so tipično izdelana za formalna besedila, napisana v standardnem jeziku (npr. novice, eseji ali pravna besedila), ne delujejo dobro nad vsebinami, ki jih ustvarjajo uporabniki (npr. tviti). V prispevku predstavimo voden način za prepoznavanje in klasifikacijo imenskih entitet v hrvaških tvitih. Primerjava treh različnih modelov za označevanje zaporedij (HMM, CRF in SVM) je pokazala, da je najboljši model za to nalogo CRF, ki doseže za mikropovprečeno mero F_1 rezultat prek 87 %. Pokažemo tudi, da najboljši model za prepoznavanje hrvaških imenskih entitet v standardnem jeziku deluje mnogo slabše kot naši modeli za prepoznavanje imenskih entitet v tvitih.

1. Introduction

Named Entity Recognition (NER) is a well-known task in information extraction (IE) and natural language processing (NLP), which aims to extract and classify names (personal names, organizations, locations), temporal expressions, and numerical expressions appearing in natural language texts. For many applications (e.g., journalism, intelligence, historical research) named entities carry the piece of information that is crucial for understanding and interpreting the text. Robust named entity recognition is also essential for other IE and NLP tasks (e.g., relation extraction and sentiment analysis). For example, to identify towards whom the sentiment is expressed in news analysis, one first needs to identify people and organizations mentioned in news stories.

NER systems typically extract named entities from documents written in standard language (e.g., news stories, essays, manuals, legal documents, police reports), i.e., documents for which the correctness of language (spelling, grammar, vocabulary) is typically checked prior to their publishing. In contrast, a lot of textual content on the web that may contain valuable information (e.g., forums, blogs, posts on social networks) is user-generated, which means that it is written in informal and colloquial language. Such language is often orthographically and grammatically incorrect, and abounds with social-media jargon. This makes user-generated text very challenging for automated processing. It has been shown (Liu et al., 2011) that the performance of the standard NER systems drops significantly when applied to informal text

In this paper we address the task of named entity extraction from tweets in Croatian. Tweets are messages from a micro-blogging service Twitter in which users post anything from news and trending events to personal information. The approach taken in this work is a supervised one: we first manually annotate tweets with named entities and then train supervised machine learning models to automatically recognize named entities in tweets. We experiment with three different supervised models – a Hidden Markov Model (HMM), Conditional Random Fields (CRF), and Support Vector Machines (SVM) – and compare their performance in a relaxed and strict evaluation settings. To the best of our knowledge, this is the first work on named entity extraction from tweets for Croatian or a Slavic language in general.

The rest of the paper is structured as follows. In the next section, we give an overview of work on NER from tweets and NER for Croatian. In Section 3, we describe the dataset and the annotation process in more detail. In Section 4, we describe the different models and features used for the task, whereas in Section 5 we present and discuss the performance for all models. Finally, we conclude and outline ideas for future work in Section 6.

2. Related work

While there is an immense body of work on named entity recognition from texts written in standard language for various languages (Finkel et al., 2005; Faruqui et al., 2010; Cucchiarelli and Velardi, 2001; Poibeau, 2003), the work on named entity extraction from tweets is rather recent and so far virtually limited to English (Finin et al., 2010; Liu et al., 2011; Ritter et al., 2011; Li et al., 2012).

Finin et al. (2010) experimented with annotating named entities in tweets in English using crowdsourcing, which showed to be rather effective, fast, and cheap. Liu et al. (2011) use a semi-supervised approach to recognize and classify named entities in English tweets. They employ k-nearest neighbors (k-NN) classifier to pre-label the tweets and sequence labeling with CRF to capture fine-grained information encoded in tweets. Ritter et al. (2011) develop a POS-tagger, a shallow parser, and a named entity recognizer for English tweets by considering both in-domain and

out-of-domain data. Their NER system exploits the output of a tweet-adjusted POS-tagger, but also employs distant supervision by applying topic modeling with constraints based on a Freebase dictionary of entities. Unlike aforementioned supervised attempts, Li et al. (2012) introduce an unsupervised, two-step NER system for targeted Twitter streams. In the first step they partition the tweets into NE candidates, which they then rank using a random-walk model based on the intrinsic properties of Twitter streams.

A number of NER systems for standard Croatian have been developed, both rule-based (Bekavac and Tadić, 2007) and statistical ones (Ljubešić et al., 2012; Karan et al., 2013). Ljubešić et al. (2012) train the Stanford NER model (Finkel et al., 2005) on Croatian data manually annotated with basic classes of named entities (PERSON, ORGANIZATION, LOCATION, MISC). Karan et al. (2013) developed CroNER, a supervised NER system using sequence labeling with conditional random fields (CRF). CroNER employs a rich set of lexical and gazetteer-based features and enforces document-level consistency of individual classification decisions. CroNER annotates nine classes of named entities and is considered to be a state-of-art NER system for Croatian (Agić and Bekavac, 2013; Karan et al., 2013).

Like CroNER, in this work we also use sequence labeling algorithms for named entity recognition and classification. However, our models are trained on manually annotated tweets instead of standard-language texts. Similarly to Ljubešić et al. (2012), we focus on three main classes of named entities: PERSON, ORGANIZATION, and LOCATION. To confirm that extracting named entities from tweets is different from extracting named entities from standard text, we evaluated CroNER on the tweets dataset, where it exhibited a significant drop in performance.

3. Dataset and annotations

In our work we use the corpus of Croatian tweets compiled by Ljubešić et al. (2014) with the open-source tool TweetCaT. TweetCaT is designed to construct Twitter corpora for smaller languages like Croatian and Slovene by collecting the URLs of web pages from seed terms. The Croatian tweet corpus contains approximately 26 million tweets. However, a fairly large portion of tweets is in Serbian language. To ease filtering, each tweet has been automatically tagged with a language identification tag. From tweets tagged as Croatian, we selected a sample 5.000 tweets for manual annotation. We subsequently removed some tweets because they were informationally irrelevant (e.g., *"Ivana Ivana Ivana Ivana"*), leaving us with the final dataset of 4.667 tweets. Further inspection revealed that roughly 30% of tweets tagged as Croatian are actually written in Serbian, and that additional manual filtering would be required to obtain a clean dataset. Because of the considerable effort involved, we decided not to perform additional filtering, but instead decided to use the corpus with mixed Croatian and Serbian tweets.[1]

To speed up the annotation process, we performed semi-automated instead of fully manual annotation. Before initiating the semi-automated annotation, we compiled the annotation guidelines, some of which adopted from Finin et al. (2010):

- Annotate each token separately, following the B-I-O annotation scheme (e.g., *Hrvatska [B-ORG] narodna [I-ORG] banka [I-ORG]*);
- Annotate names, surnames, and nicknames but not their titles (e.g., *doc. dr. sc.* as instances of the PERSON class (e.g., *Marko [B-PER]*; *dr. Ivo [B-PER] Josipović [I-PER]*);
- Annotate names of concrete organizations, institutions, state authorities, sport clubs, national teams, but not generic terms like *government* or *party* as instances of the ORGANIZATION class (e.g., *NK [B-ORG] Rijeka [I-ORG]*);
- Annotate mentions of places, regions, states, rivers, mountains, squares, streets, etc. as instances of the LOCATION class (e.g., *Velika [B-LOC] Gorica [I-LOC]*);
- Do not annotate tokens starting with "@";
- Do annotate named entities preceded by "#";
- Annotate words according to the tweet context (e.g., token *"Rijeka"* may denote the location but it may also be part of the organization mention *"NK Rijeka"*);
- When in doubt whether to annotate the word as an instance of LOCATION or ORGANIZATION class, prefer ORGANIZATION.

Semi-automated annotation. The semi-automated annotation consists of two steps: (1) automated annotation of all mentions found in precompiled gazetteers and (2) manual correction of errors (both false positives and false negatives) made by the automated gazetteer-based annotation. This automated gazetteer-based annotation was also used as a baseline for the evaluation of supervised models. To perform the first step of the semi-automated annotation, we first needed to compile the set of gazetteers. Gazetteers with personal names (2413 entries) and locations (71 entries) were obtained from individual web resources.[2,3] A gazetteer with organization names (109 entries) was compiled from several different web resources. Following the automated gazetteer-based annotation, we manually corrected all errors introduced by the automated annotator. We also labeled named entity mentions omitted by the automated annotator. Organization mentions were most frequently omitted by the automated annotator because of (1) the limited size of organizations gazetteer and (2) the fact that the organizations gazetteer contained only single-word entries and organizational mentions quite often consists several words.

[1] Arguably, from a machine learning perspective, using a mixed Croatian-Serbian corpus as the train set introduces some noise in all cases in which the differences between the two languages are reflected in the feature values. On the other hand, our preliminary experiments, carried out on separate Croatian and Serbian test sets, have shown that the model performs equally well on both test sets. Thus, the upside of using a noisy dataset in this case is that one gets a model that works reasonably well for both languages.

[2] http://www.croatian-genealogy.com

[3] http://goo.gl/79ddLr

Class	MUC F_1 (%)	Exact F_1 (%)
PERSON	94.7	92.8
ORGANIZATION	85.7	81.2
LOCATION	86.6	85.2
Micro-average	91.3	88.8

Table 1: Inter-annotator agreement.

Many locations were also omitted because only names of Croatian cities were in the location gazetteer. Person names were omitted rather rarely, primarily due to the size of the corresponding gazetteer.

Manual annotation. The manual annotation step was performed by two annotators (the first two authors). Initially, both annotators independently annotated the same set of 500 tweets to measure the inter-annotator agreement (IAA) and assess how well the annotation guidelines are followed. The IAA was measured by computing both MUC and Exact F_1-scores between the annotations of the two annotators. In the MUC scheme two annotations are considered the same if they have the same class and their extents overlap in at least one token. In the Exact evaluation scheme, the match is only counted when the two annotation are exactly the same (same class and exactly the same extent). IAA scores for all NE classes are given in Table 1. After annotating the same initial 500 tweets, each of the annotators annotated a separate set of approximately $2,230$ tweets. These tweets were used for training and testing the models.[4]

4. NER models

4.1. Machine learning models

We used three different supervised machine learning models to extract and classify named entities in tweets: (1) a Hidden Markov Model (HMM), (2) Conditional Random Fields (CRF), and (3) a Supported Vector Machine (SVM). For all three models, we used the implementation in NLTK,[5] a popular Python library for natural language processing.

Hidden Markov Model. This model is an extension of Markov process where each state has all observations joined by the probability of the current state generating observation (Blunsom, 2004). Formally, HMM is defined as a tuple:

$$HMM = \big(S, O, A, B, \pi\big), \tag{1}$$

where *S* denotes hidden states (in our case labels of tokens), *O* stands for outputs in each state (in our case all words seen in tweets) and three parameters that denote probabilities computed from the annotated corpus: the starting probability π, transition probabilities A of going from one state to the other, and output probabilities *B*, in other words the probability of seeing a word when in one particular state.

Support Vector Machines. The standard SVM is a binary classification algorithm, which performs classification by maximizing the margin between the examples of the two classes. The binary SVM formulation can be easily extended to account for multi-class classification problems. However, in this work we employ a structured, sequence labeling variant of the SVM, proposed by Altun et al. (2003). Sequence labeling formulation of the SVM is very similar to the multi-class SVM formulation with exponentially many classes.

Conditional Random Fields. CRF is a discriminative probabilistic graphical model that can model overlapping, non-independent features in a sequence of data. A special case, linear-chain CRF, can be thought of as the *undirected graphical model* version of the HMM. Unlike HMM, CRF allows to extract arbitrary features for the current token as well as for preceding and following tokens. We used a window of size five for extracting the features, i.e., all of the features were computed for the current token and the two tokens preceding and succeeding it.

4.2. Features

Due to the nature of the models, slightly different feature sets were used for each of them. The following list is the union of the features used for all three models:

- f^1 – The lemma of the token;
- f^2 – The length of the token;
- f^3 – The shape of the token encodes the lower/upper casing of the word (e.g., the shape of the word *Ana* is ULL);
- f^4 – A feature indicating whether the token contains a non-alphanumeric character (e.g., *Lovrić-Merzel*);
- f^5 – A feature indicating whether the token contains only non-alphanumeric characters (e.g., *?!*);
- f^6 – Features indicating whether the token is the first or the last token in the tweet
- f^7 – A feature indicating whether the token contains any lower-cased letters;
- f^8 – A feature indicating whether the token contains any upper-cased letters;
- f^9 – A feature indicating whether the token contains any alphanumeric characters;
- f^{10} – A feature indicating whether the token contains digits (e.g., *sk8*);
- f^{11} – Features indicating whether the token matches a gazetteer entry (one feature per gazetteer, as a token can match multiple gazetteer entries).

For HMM, we used only one feature - the lemma of the word (f^1) – as other features cannot be incorporated into the standard HMM model. For the other two models – CRF and structured SVM – we used all above-mentioned features (f^1–f^{11}).

[4]The annotated dataset is available under the Creative Commons BY-NC-SA license from
http://takelab.fer.hr/cronertweet

[5]http://www.nltk.org/

	Baseline			HMM			SVM			CRF		
NE class	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
PERSON	96.47	84.38	90.02	93.83	81.46	87.21	90.74	88.25	89.46	94.83	92.72	93.76
LOCATION	50.00	27.30	35.32	90.00	16.02	27.20	52.16	39.47	44.93	78.35	68.77	70.33
ORGANIZATION	74.26	45.56	56.48	87.64	45.86	60.22	73.33	44.66	55.51	76.94	75.80	76.37
Overall macro	73.58	52.42	60.60	**90.49**	47.78	58.21	72.08	57.46	63.31	83.37	**79.10**	**81.13**
Overall micro	88.38	68.37	77.10	**92.63**	65.21	76.54	83.77	72.11	77.50	89.01	**86.10**	**87.53**

Table 2: MUC evaluation results.

	Baseline			HMM			SVM			CRF		
NE class	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
PERSON	64.48	55.90	59.89	84.64	73.90	78.90	82.22	80.91	81.56	89.18	88.00	88.58
LOCATION	46.20	25.22	32.63	86.67	15.43	26.20	50.20	37.98	43.24	71.65	62.90	66.99
ORGANIZATION	38.13	23.91	29.39	69.50	35.64	47.12	55.80	34.74	42.82	66.08	65.43	65.76
Overall macro	49.60	35.01	40.64	**80.27**	41.66	50.74	62.74	51.21	55.87	75.64	**72.11**	**73.78**
Overall micro	57.87	44.67	50.42	**82.10**	57.85	67.88	74.43	64.85	69.31	82.09	**79.99**	**81.03**

Table 3: Exact evaluation results.

5. Evaluation

5.1. Experimental setup

We split the tweets dataset into two or three sets, depending on the learning algorithm. Since HMM only uses lemmas as features, we did not have to perform feature selection as for the other two algorithms. Thus, for HMM we split the tweets into two sets: train set (3399 tweets) and test set (1268) tweets. We trained HMM on the train set and we report the performance of the model on the test set. For SVM and CRF we performed greedy backward feature selection to identify the best subset of features for the task. Thus, we split the dataset into three subsets: train set (3399 tweets), validation set (423 tweets), and test set (845) tweets. For both algorithms we optimized the set of features according to the performance on the validation set. We report the performance for CRF and SVM with optimal feature subsets on the test set. As the baseline we used the same automated method that we employed as the first step of the semi-automated annotation process – the token is tagged as a named entity of some type if it can be found in the gazetteer for that NE type. Additionally, the baseline merges adjacent tokens found in the same gazetteer into a single named entity mention.

5.2. Results

The performance for all three models and the baseline, measured for MUC and Exact setting, is given in Table 2 and Table 3, respectively. The performance is reported for each of the NE classes, along with both micro-averaged and macro-averaged performance.

The CRF model outperforms the other two models by a wide margin in both evaluation settings. This is the consequence of CRF taking into account features of the preceding and following tokens as well. Thus, it is able to learn the patterns of named entity occurrence much better then the other models. Interestingly, HMM exhibits best precision but very low recall in both evaluation settings. In the MUC setting, HMM model does not even outperform the baseline in terms of F_1-score.

The structured SVM consistently outperforms the baseline and the HMM model, but is also consistently outperformed by the CRF model. The most common cause of errors for the structured SVM model are tokens labeled as inside of a named entity (e.g., I-PER) even when the preceding token was not the beginning of a named entity (e.g., B-PER). In contrast, CRF assigns very low probabilities for the "I-" labels when previous label in the sequence is not "B-".

To asses the performance of the NER system built for texts written in standard language, we evaluated CroNER (Karan et al., 2013) on the test portion of the annotated Twitter dataset. The results for Croatian are in line with the observations for English (Liu et al., 2011) – the performance of the tagger built for texts written in standard language drops significantly when applied to tweets. CroNER exhibited micro-averaged performance of 35.8% F_1-score in the MUC setting, and merely 27.4% F_1-score of in the Exact evaluation setting.

6. Conclusion

Traditional IE and NLP tools have been shown ineffective when applied to user-generated content. This is especially true for tweets, micro-blogging messages filled with jargon vocabulary and abbreviations. In this paper we presented the work on named entity recognition for Croatian tweets. We semi-automatically annotated the collection of almost 5.000 tweets in Croatian and Serbian. We compared three different sequence labeling models, demonstrating that CRF, being able to incorporate context features, outperforms HMM and structured SVM as well as the gazetteer-based baseline. The overall performance of the CRF model (87% micro-averaged MUC F_1-score) is comparable to the performance of the state-of-the-art NER system for Croatian

standard language (90% micro-averaged MUC F_1-score; Karan et al. (2013)), which we consider very encouraging considering the lack of POS and syntactic information in current models. We also demonstrated that a NER system built for standard language texts performs poorly on tweets.

There are several possible extensions of the work presented in this paper. Firstly, we intend to extend the models with part-of-speech and syntactic information. This means that a designated POS-tagger and (shallow) parser for tweets need to be created for Croatian and Serbian as, similar to NER, standard tools have been shown inefficient. Secondly, a Twitter dataset could be enlarged in order to determine how the dataset size influences the performance of the tagger. Finally, we believe that enforcing consistency of named entity annotations across tweets of the same thread (re-tweets) would improve the overall performance.

7. References

Ž. Agić and B. Bekavac. 2013. Domain-aware evaluation of named entity recognition systems for Croatian. *CIT. Journal of Computing and Information Technology*, 21(3):185–199.

Y. Altun, I. Tsochantaridis, T. Hofmann, et al. 2003. Hidden markov support vector machines. In *ICML*, volume 3, pages 3–10.

B. Bekavac and M. Tadić. 2007. Implementation of Croatian NERC system. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, pages 11–18. Association for Computational Linguistics.

P. Blunsom. 2004. Hidden markov models. *Lecture notes, August*, 15:18–19.

A. Cucchiarelli and P. Velardi. 2001. Unsupervised named entity recognition using syntactic and semantic contextual evidence. *Computational Linguistics*, 27(1):123–131.

M. Faruqui, S. Padó, and M. Sprachverarbeitung. 2010. Training and evaluating a german named entity recognizer with semantic generalization. In *Proc. of KONVENS*, pages 129–133.

T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. 2010. Annotating named entities in Twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88. Association for Computational Linguistics.

J. R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

M. Karan, G. Glavaš, F. Šarić, J. Šnajder, J. Mijić, A. Silić, and B. D. Bašić. 2013. CroNER: recognizing named entities in Croatian using conditional random fields. *Informatica (Slovenia)*, 37(2):165–172.

C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee. 2012. Twiner: named entity recognition in targeted Twitter stream. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 721–730. ACM.

X. Liu, S. Zhang, F. Wei, and M. Zhou. 2011. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 359–367. Association for Computational Linguistics.

N. Ljubešić, M. Stupar, and T. Jurić. 2012. Building named entity recognition models for Croatian and Slovene. In *Proceedings of the Eighth Information Society Language Technologies Conference*, pages 117–122.

N. Ljubešić, D. Fišer, and T. Erjavec. 2014. TweetCaT: a tool for building Twitter corpora of smaller languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

T. Poibeau. 2003. The multilingual named entity recognition framework. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2*, pages 155–158. Association for Computational Linguistics.

A. Ritter, S. Clark, O. Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics.

Discriminating between VERY similar languages among Twitter users

Nikola Ljubešić, Denis Kranjčić

Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences
University of Zagreb
I. Lučića 3, HR-10000 Zagreb
{nljubesi,dkranjci}@ffzg.hr

Abstract

In this paper we tackle the problem of discriminating Twitter users by the language they tweet in, taking into account very similar South-Slavic languages, namely Bosnian, Croatian, Montenegrin and Serbian. We take the supervised machine learning approach by annotating a subset of 500 users from an existing Twitter collection by the language they primarily tweet in. We show that by using either words or character 6-grams as features, univariate feature selection, up to 10% of most significant features and a standard classifier, on the user level we reach accuracy of ∼97%.

Razlikovanje med ZELO podobnimi jeziki uporabnikov Twitterja

V prispevku raziskujemo problem ločevanja uporabnikov družabnega omrežja Twitter glede na to, v katerem jeziku tvitajo, pri čemer obravnavamo zelo podobne južnoslovanske jezike: bosanščino, hrvaščino, srbščino in črnogorščino. Uporabimo pristop z nadzorovanim strojnim učenjem, kjer označimo vsakega uporabnika iz že obstoječe podatkovne množice 500 uporabnikov z jezikom, v katerem tvita. Pokažemo, da z uporabo besed ali 6-gramov znakov kot značilk, univariantno izbiro značilk, do 10% najpomembnejših značilk in standardnega klasifikatorja dosežemo ∼97% točnost pravilne klasifikacije posameznega uporabnika.

1. Introduction

The problem of language identification, which was considered a solved task for some time now, has recently gained in popularity among researchers by identifying more complex problems such as discriminating between language varieties (very similar languages and dialects), identifying languages in multi-language documents, code-switching (alternating between two or more languages) and identifying language in very short documents (such as tweets).

In this paper we address the first and the last problem, namely discriminating between very similar languages in Twitter posts, with the restriction that we do not identify language on the tweet level, but the user level.

The four languages we focus on here, namely Bosnian, Croatian, Montenegrin and Serbian, belong to the South Slavic group of languages and are all very similar to each other.

All the languages, except Montenegrin, use the same phonemic inventory, and they are all based on the write-as-you-speak principle. Croatian is slightly different in this respect, because it does not transcribe foreign words and proper nouns, as the others do. Moreover, due to the fairly recent standardization of Montenegrin, its additional phonemes are extremely rarely represented in writing, especially in informal usage. The Serbian language is the only one where both Ekavian and Ijekavian pronunciation and writing are standardized and widely used, while all the other languages use Ijekavian variants as a standard. The languages share a great deal of the same vocabulary, and some words differ only in a single phoneme, because of phonological, morphological and etymological circumstances. There are some grammatical differences regarding phonology, morphology and syntax, but they are arguably scarce and they barely influence mutual intelligibility. The distinction between the four languages is based on the grounds of establishing a national identity, rather than on prominently different linguistic features.

2. Related work

One of the first studies incorporating similar languages in a language identification setting was that of Padró and Padró (2004) who, among others, discriminate between Spanish and Catalan with an accuracy of up to 99% by using second order character-level Markov models. In (Ranaivo-Malancon, 2006) a semi-supervised model is presented to distinguish between Indonesian and Malay by using frequency and rank of character trigrams derived from the most frequent words in each language, lists of exclusive words, and the format of numbers. Huang and Lee (2008) use a bag-of-words approach to classify Chinese texts from the mainland and Taiwan with results of up to 92% accuracy. Zampieri and Gebre (2012) propose a log-likelihood estimation method along with Laplace smoothing to identify two varieties of Portuguese (Brazilian and European) obtaining 99.5% accuracy.

In the first attempt at discriminating between the two most distant out of the four languages of interest, namely Croatian and Serbian, Ljubešić et al. (2007)have shown that by using a second-order character Markov chain and a list of forbidden words, the two languages can be differentiated with very high accuracy of $\sim 99\%$. As a follow-up, Tiedemann and Ljubešić (2012) add Bosnian to the language list showing that most off-the-shelf tools are in no way capable of solving that problem, while their approach by identifying blacklisted words, reaches accuracy of ∼97%. Ljubešić and Klubička (2014) have worked with the same three languages as a subtask of producing web corpora of those languages. They have shown to outperform the best performing classifier from (Tiedemann and

Ljubešić, 2012) by training unigram language models on the whole content of the collected web corpora showing to decrease the error on the Croatian–Serbian language pair to a fourth. Recently, as part of the DSL (Discriminating between Similar Languages) 2014 shared task on discriminating between six groups of similar languages on the sentence level (Zampieri et al., 2014), the language group A consisted of Bosnian, Croatian and Serbian and the best result in the group yielded 93.6% accuracy, which is not directly comparable to the previously reported results because classification was performed on sentence level, and not on document level as in previous research.

To best of our knowledge, there has been only one attempt at discriminating between languages of that level of similarity, namely Croatian and Serbian, on Twitter data in (Ljubešić et al., 2014) where word unigram language models built from the Croatian and Serbian web corpora were used in the attempt at separating users by those two languages. An analysis of the annotation results showed that there is a substantial Twitter activity of speakers of both Bosnian and Montenegrin and that the the collected data cannot be described with the two-language classification schema, but with a 4-class schema which takes into account all the languages in the collection.

Our work builds on top of this previous research by defining a four-language classification schema, inside which Montenegrin, a language that gained official status in 2007, is present for the first time. Additionally, this is the first focused attempt on discriminating between those languages – and possibly between such similar languages overall – on Twitter data.

3. Dataset

The dataset we run our experiments on consists of tweets of 500 random users from the Twitter collection obtained with the TweetCat tool described in (Ljubešić et al., 2014).

There was only one annotator available for this annotation task. Annotating a portion of the dataset by multiple users is considered future work.

Having other languages in the dataset (mostly English) was tolerated as long as most of the text was written in the chosen language. Beside the four main categories, one user, tweeting in Bosnian, had most of the tweets in English (preprocessing error), there was one user tweeting in Macedonian and 8 users were tweeting in Serbian, but using the Cyrillic script. Those 10 users were discarded from the dataset and the following experiments. The users tweeting in Serbian and using the Cyrillic scripts were discarded because we want to concentrate here on discriminating between the languages based on content and not the script used.

The result of the annotation procedure is summarized in the distribution of users given their language presented in Table 1. We can observe that Serbian makes up 77% of the dataset, that there is a similar amount, around 9%, of Bosnian and Croatian data, while Montenegrin is least represented with around 5% of the data. These results are somewhat surprising because there is a much higher number of speakers of Croatian (around 5 million) than of Bosnian (around 2 million) or Montenegrin (below 1 million).

language (code)	# of users
Bosnian (bs)	46
Croatian (hr)	42
Montenegrin (me)	24
Serbian (sr)	378

Table 1: Distribution of users by the language they tweet in

	token	3-gram	6-gram
GNB	0.788	0.769	0.780
KNN	0.780	0.771	0.786
DT	0.894	0.892	0.871
SVM	0.881	0.887	0.897
RF	0.839	0.835	0.843
AB	0.861	0.869	0.876

Table 2: Obtained accuracies in the initial experiments with different classifiers and features

4. Experiments

We perform data preprocessing, feature extraction and data formatting to the svmlight format with simple Python scripts. All the experiments are carried out with the machine learning kit scikit-learn (Pedregosa et al., 2011). Our evaluation metric is accuracy calculated via stratified 5-fold cross-validation.

Each instance in our experiments is one of the 490 annotated Twitter users. We extract features only from the preprocessed text of each user. We could use the information about each specific user like their name, bio, location etc., but we leave this line of research for future work. During preprocessing we remove URLs, hashtags and mentions from the text of each user as well. By preparing our dataset in the described fashion, we remove all the specificities of Twitter generalizing to any sort of user-generated content.

After performing preprocessing, the average number of words per user is 6,606.53 words, with a minimum of 561 and a maximum of 29,246 words.

4.1. Initial experiment

The aim of the initial experiment was to get a feeling for the problem at hand by experimenting with various classifiers and features.

We experiment with the traditional classifiers, such as the Gaussian naive Bayes (GNB), k-nearest neighbor (KNN), decision tree (DT) and linear support-vector machine (SVM), as well as classifier ensembles such as AdaBoost (AB) and random forests (RF). For each classifier we use the default hyperparameter values except for the linear SVM classifier for which we do tune the C parameter for highest accuracy.

From previous research we know that best features for discriminating between similar languages are words and longer character n-grams (around level 6). Traditionally, in the task of language identification, character 3-grams were

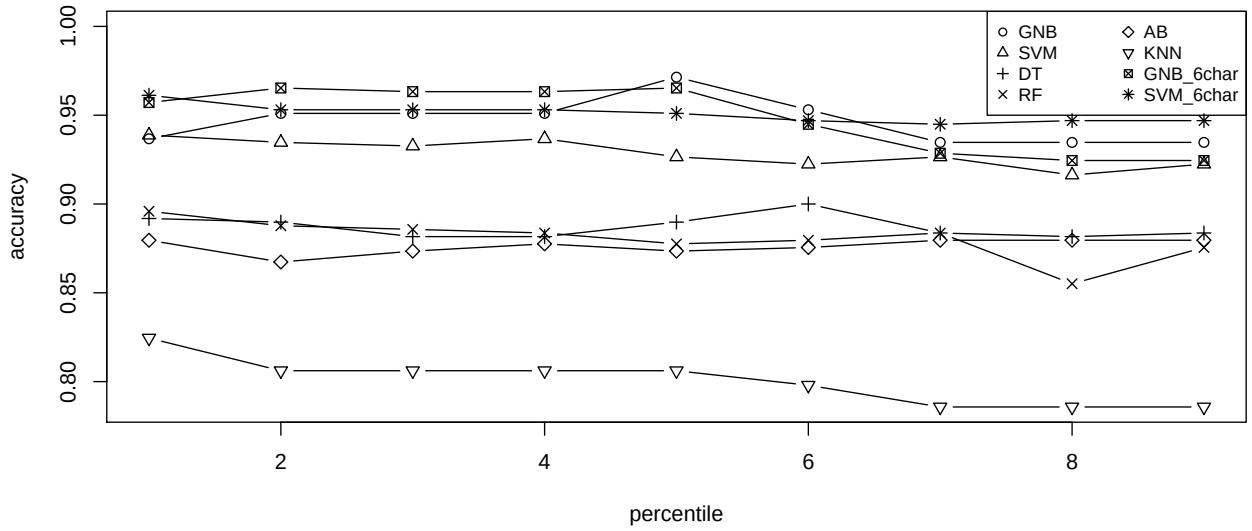


Figure 1: Accuracy of each classifier given the percentile of features with minimal p-values used

most frequently used. This is why we run our initial experiments with three sets of features: tokens, character 3-grams and character 6-grams. We extract character n-grams from tokens with one special character added to the beginning and end of the token. While extracting 6-grams, we add tokens shorter than 4 characters (6 characters with the surrounding special characters) to the feature set as well.

We compare the classifiers by calculating accuracy on 5-fold stratified cross-validation. The results are given in Table 2. We can observe that each set of features produces very similar results, the character 3-gram underperforming slightly, and that the differences between the results are due to usage of a specific classifier. DT and SVM obtain best results while GNB and KNN perform the worst, just slightly above the most-frequent-class baseline. The low score of the GNB classifier, which has no inherent feature selection, and the overall best results obtained by the simple DT classifier, which has implicit feature selection, hint that our results could improve if we applied explicit feature selection as a pre-processing step. This follows our intuition that similar languages can be discriminated through a limited number of features and not the whole lexicon or character n-gram set.

We continue our experiments by introducing a feature selection algorithm and using tokens as our 213,246 initial feature list because of their easier interpretability.

4.2. Feature selection

Although there are stronger feature selection algorithms, we opt for the simple univariate feature selection algorithm which sorts features by their p-value through the F1 ANOVA statistical test and chooses the user-specified percentile of features from the bottom of the list. We use this simple feature selection method because we assume independence of our features, i.e. tokens or character n-grams, which mostly stands in the problem of language identification. Here we experiment with all the classifiers from the previous subsection and the percentile of strongest features ranging from 1 to 9 since all classifiers reach their best performance in that range. The results are shown in Figure 1.

The two best-performing classifiers, once the number of features is down to single-digit percentiles, are the GNB and the SVM. The overall best performing setting is the GNB, which uses 5 percentiles of features (0.971). The worst performing classifier is the KNN which yields worse performance as the number of features increases.

We did perform experiments with other feature sets as well, obtaining very similar results when using character 6-grams (GNB peaking at 2 percentiles with 0.965 and SVM peaking at 1 percentile with 0.961, both depicted in Figure 1) and obtaining worse results when using character 3-grams (0.816 with GNB on 13 percentiles of features and 0.945 with SVM on 4 percentiles of features). Combining the character 6-gram and token feature sets did not produce any improvements, which is to be expected because those two feature sets contain very similar information.

We can consistently observe the phenomenon that SVM outperforms GNB on smaller number of features and on features of lower quality. Although these properties can be important if speed and memory consumption are of great importance, or if no better features are at our disposal, here we choose the GNB on 5 percentiles of features as our final classifier because of its exceedingly superior accuracy. By using 5 percentiles of features, we shrink our model from the initial 213,246 features down to 10,662.

4.3. Confusion matrix and strongest features

We take a closer look at our best performing classifier by plotting our confusion matrix and by calculating preci-

sion and recall on each class. The plot is given in Table 4. We can observe that the two most problematic languages are Bosnian being confused with Serbian and Croatian, and Montenegrin being confused with Serbian.

Next, we inspect the most informative 50 features from our feature selection algorithm and present them, along with the a-posteriori parameter values for each language, in Table 3. While there are a few features that are concept-oriented and not language-specific, such as the toponyms Zagreb and Podgorica (the capitals of Croatia and Montenegro), most features are language-specific and of possible interest to linguists. This is why we will publish all the selected features with the corresponding parameter values for all four languages.

4.4. Learning curve

Finally, we compare our two best-performing classifiers, GNB and SVM by plotting learning curves, using the best performing percentile of features for each classifier. The learning curves are depicted in Figure 2 showing that GNB does outperform SVM on all training data sizes and that there is still room for improvement by moderately increasing the amount of available data.

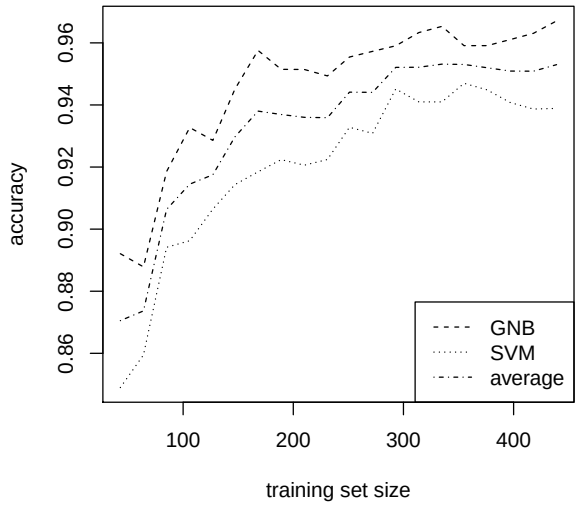


Figure 2: Learning curves of the GNB and SVM classifiers after feature selection

5. Error analysis

We performed error analysis by reinspecting tweets of users that were differently classified by the best performing automated classifier and the human annotator.

We identified altogether 5 users that were incorrectly classified by the human annotator because the bulk of their tweets consisted of retweets and tweets written in languages such as English and German. In those cases, original tweets in the users' native language were very scarce, which made the manual annotation very tiresome. The fact that a third of assumably wrongly classified users are actually human

feature	bs	hr	me	sr
sjutra	0.065	0.048	4.708	0.013
prije	3.152	4.548	5.042	0.119
vrijeme	3.543	5.214	5.292	0.146
dje	1.022	0.0	5.083	0.143
mjesta	0.435	1.19	0.667	0.011
podgorice	0.043	0.0	0.833	0.013
uvijek	4.826	5.69	4.125	0.164
točno	0.022	0.667	0.0	0.003
dio	0.609	1.905	0.625	0.032
cijeli	1.13	1.167	1.292	0.016
poslije	1.87	0.69	1.792	0.048
netko	0.022	2.762	0.0	0.034
gdje	4.0	3.786	1.792	0.101
pg	0.13	0.0	1.875	0.013
tko	0.152	5.357	0.167	0.053
sretan	1.0	2.595	0.042	0.071
podgorica	0.022	0.0	2.0	0.029
cus	0.0	0.0	0.625	0.0
mjesto	0.522	1.119	1.0	0.029
mjestu	0.696	0.571	0.5	0.011
mjeseca	0.391	0.762	0.875	0.008
vjerujem	1.239	0.548	1.042	0.034
lijepo	1.652	2.19	1.5	0.053
zagrebu	0.109	1.643	0.042	0.09
dvije	1.239	1.357	2.5	0.058
ovdje	2.043	1.619	1.0	0.026
podgorici	0.0	0.095	1.0	0.034
vjerovatno	0.5	0.071	0.708	0.005
mjesec	0.891	1.119	1.417	0.045
tjedan	0.0	2.238	0.0	0.003
kuna	0.0	0.881	0.0	0.003
podgoricu	0.0	0.024	0.5	0.008
lijep	0.674	0.595	0.667	0.019
dako	0.022	0.0	0.333	0.0
kava	0.043	1.0	0.042	0.011
bit	0.848	3.857	2.167	0.198
vjerojatno	0.022	0.69	0.0	0.0
ljeto	0.696	1.048	1.75	0.045
pjesme	1.0	0.762	1.333	0.063
umjesto	1.152	0.905	1.167	0.053
kruh	0.0	0.238	0.0	0.0
cg	0.152	0.0	2.625	0.146
zagreb	0.109	2.31	0.0	0.037
svatko	0.0	0.524	0.0	0.011
vidjeti	0.435	0.857	0.292	0.024
negdje	1.13	0.762	0.708	0.019
vazda	0.326	0.0	1.708	0.071
zabolje	0.0	0.0	0.333	0.0
vidji	0.0	0.0	0.375	0.005
pjesma	0.957	0.714	1.25	0.04
djevojkama	0.109	0.024	0.458	0.003

Table 3: 50 strongest features by the univariate feature selection algorithm with per-language parameter values from the GNB classifier

	bs	hr	me	sr	P	R
bs	44	0	0	2	0.917	0.957
hr	1	41	0	0	0.976	0.976
me	0	0	21	3	0.850	0.875
sr	3	1	4	370	0.987	0.979

Table 4: Confusion matrix along with precision and recall for the best performing classifier

annotator errors has motivated us further in including additional annotators in the future.

The remaining 9 manually correctly annotated users were partially wrongly classified because of retweeting. The register in which the users tweet also affected the classification at times. For example, in the almost exclusively colloquial and informal Montenegrin part of the dataset, the only user (a news agency) who tweeted in a more formal register was wrongly classified as belonging to the more inclusive Serbian part of the dataset. It has also been noticed that some users use several languages from the classification schema throughout their tweets, in form of citations and song lyrics. Mixing of these four languages is possible in many contexts, so a dose of indecisiveness in their classification should not be surprising. For that reason we will label each user in our collection not only by the most probable language, but with the distribution of probabilities for all four languages.

6. Conclusion and future work

We have presented a supervised approach to discriminating between very similar languages on Twitter data by classifying each user to the language he or she uses predominantly.

We have annotated 500 users by their predominant language and used that data for experimenting via cross-validation. By using textual features only, we have shown that very similar performance is obtained when using character n-grams or tokens as features. We have shown that feature selection significantly improves the results, which is to be expected given the problem at hand. We obtained very similar results when using linear SVM or Gaussian NB, linear SVM performing better on smaller sets of features or less informative features like character 3-grams, but overall best performance of 97.1% accuracy was obtained using 5% of features and Gaussian NB.

The worst performing language was Montenegrin, being quite often mixed with Serbian, and the second worst Bosnian, being mixed with both Serbian and Croatian.

Next steps include annotating the sample by multiple users for obtaining inter-annotator agreement rates and improving accuracy, as the learning curves suggest. Additionally, at this point only the text of the tweets was used and usage of additional features such as geo-location and user profile information should be inspected as well.

We release the annotated Twitter user lists as well as the prepared datasets in the svmlight format[1] under the CC-BY-SA 4.0 license[2].

7. References

Chu-Ren Huang and Lung-Hao Lee. 2008. Contrastive approach towards text source classification based on top-bag-of-word similarity. In *PACLIC*, pages 404–410. De La Salle University (DLSU), Manila, Philippines.

Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden. Association for Computational Linguistics.

Nikola Ljubešić, Nives Mikelić, and Damir Boras. 2007. Language identification: How to distinguish similar languages. In Vesna Lužar-Stifter and Vesna Hljuz Dobrić, editors, *Proceedings of the 29th International Conference on Information Technology Interfaces*, pages 541–546, Zagreb. SRCE University Computing Centre.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2014. TweetCaT: a Tool for Building Twitter Corpora of Smaller Languages. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Lluís Padró and Muntsa Padró. 2004. Comparing methods for language identification. *Procesamiento del Lenguaje Natural*, 33:155–162, September.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Bali Ranaivo-Malancon. 2006. Automatic Identification of Close Languages – Case study: Malay and Indonesian. *ECTI Transactions on Computer and Information Technology*, 2(2):126–134.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*, pages 2619–2634, Mumbai, India.

Marcos Zampieri and Binyam Gebrekidan Gebre. 2012. Automatic identification of language varieties: The case of Portuguese. In *Proceedings of KONVENS2012 - The 11th Conference on Natural Language Processing*.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A Report on the DSL Shared Task 2014. In *Proceedings of the VARDIAL workshop*.

[1] http://nlp.ffzg.hr/data/publications/nljubesi/ljubesic14-discriminating/

[2] https://creativecommons.org/licenses/by-sa/4.0/

Predicting Croatian Phrase Sentiment Using a Deep Matrix-Vector Model

Siniša Biđin, Jan Šnajder, Goran Glavaš

University of Zagreb, Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
sinisa@bidin.cc, {jan.snajder, goran.glavas}@fer.hr

Abstract

Many sentiment analysis tasks rely on the existence of a sentiment lexicon. Such lexicons, however, typically contain single words annotated with prior sentiment. Problems arise when trying to model the sentiment of multiword phrases such as *"very good"* or *"not bad"*. In this paper, we use a recently proposed deep neural network model to classify the sentiment of phrases in Croatian. The experimental results suggest that reasonable classification of phrase-level sentiment for Croatian is achievable with such a model, reaching a performance comparable to that of an analogous model for English.

Napovedovanje sentimenta besednih zvez v hrvaščini z uporabo globinskega modela matrik vektorjev

Napovedovanje sentimenta besednih zvez v hrvaščini z uporabo globinskega modela matrik vektorjev Mnogo analiz sentimenta se zanaša na obstoj leksikona z informacijami o sentimentu. Vendar takšni leksikoni tipično vsebujejo samo posamezne besede, označene z vnaprejšnjim sentimentom. Problemi se pojavijo, ko bi želeli modelirati sentiment večbesednih enot, kot so »zelo dobro« ali »ni slabo«. V prispevku uporabimo pred kratkim predlagano globinsko nevronsko mrežo, s katero klasificiramo sentiment besednih zvez v hrvaščini. Eksperimentalni rezultati nakazujejo, da je s takim modelom mogoče doseči razmeroma dobro klasifikacijo besednih zvez glede na njihov sentiment, saj je delovanje modela primerljivo z analognim modelom za angleški jezik.

1. Introduction

The sentiment of a word, a phrase, or a document refers to its subjective attitude, polarity, or expression of feeling. The phrase *"nicely done"* has a positive, whereas *"horribly wrong"* has a negative sentiment. Sentiment analysis explores the ways of identifying or extracting sentiment from text. Applying methods of sentiment analysis on larger amounts of text, nowadays widely available on the web, allows us to do things such as attempt to judge the popularity of a product or predict the outcome of an election.

In this paper, we focus on classifying the sentiment of Croatian phrases consisting of two words. Given sentiment-labeled phrases such as *"very bad"*, *"not bad"*, and *"very good"*, we aim to train a model to correctly learn that *"bad"* bears a negative sentiment, and *"good"* a positive one. Also, the model should learn that *"very"* is an intensifier: it amplifies the sentiment of a word it is paired with. Likewise, *"not"* should be recognized as a negator, a word that inverts the sentiment of the word or a phrase it appears next to.

To learn the sentiment of Croatian bigrams, we employ a deep neural network model proposed by Socher et al. (2012). This model has shown to have good results when applied to the English language, which is something we aim to replicate for Croatian. We train and evaluate the deep neural model on two datasets of phrases, achieving performance comparable to the results obtained for English phrases.

2. Related work

This work is most closely related to two prominent areas of natural language processing: sentiment analysis and compositionality in vector spaces. Compositionality in vector spaces refers to the problem of learning a useful representation of a composition of multiple vector representations.

Focusing on compositionality, the model we use (Socher et al., 2012) is a generalization of earlier models. One model proposes vector composition through additive and multiplicative functions (Mitchell and Lapata, 2010), while another captures compositionality of words by linear combinations of nouns represented as vectors and adjectives as matrices (Baroni and Zamparelli, 2010). Finally, a general approach for sentiment analysis of phrases was laid out by Yessenalina and Cardie (2011), interesting also in that it introduces a model that uses matrices to represent words and matrix multiplication to compose them.

Another related work focusing also on sentiment analysis is the one by Socher et al. (2011), where predictions of sentence-level sentiment distributions are made using a recursive model that attempts to model sentiment via compositional semantics. Later models improve on this and achieve state-of-the-art results for the tasks of sentence-level sentiment classification (Socher et al., 2012; Socher et al., 2013), the first of which is the very model we are using here.

3. Training the matrix-vector model

To classify the phrase sentiment, we use the MV-RNN model proposed by Socher et al. (2012). This model can be applied by recursive operators to any n-gram, but we simplify it to the point where it only handles bigrams. The MV recursive neural network model derives its name from the matrix-vector representation of words. In essence, this means that each word w of a lexicon is modeled using two separate pieces of data: an n-dimensional vector $\mathbf{x}$ representing some semantic property of the word (such as sentiment) and an n-by-n matrix $\mathbf{X}$ representing the way the word influences the same semantic property of other words with which it constitutes a phrase.

$$w = (\mathbf{x}, \mathbf{X}), \quad \mathbf{x} \in \mathbb{R}^n, \quad \mathbf{X} \in \mathbb{R}^{n \times n}$$

Given an initial set of word MV-representations and some initial shared weights $\mathbf{W}$, all initialized to some (e.g.,

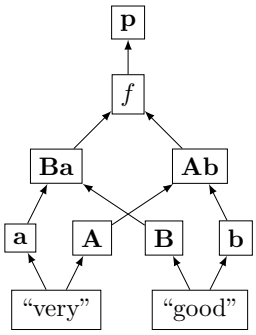


Figure 1: Two words, *"very"* and *"good"*, having MV-representations $(\mathbf{a}, \mathbf{A})$ and $(\mathbf{b}, \mathbf{B})$ respectively, affect each others' meaning (via $\mathbf{Ba}$ and $\mathbf{Ab}$) and combine using f to form a basis for phrase sentiment classification $\mathbf{p}$.

random) continuous values, in addition to a non-linear function g (e.g., a sigmoid), we can use a combining function f to determine the vector representation $\mathbf{p}$ of an entire phrase. This is depicted in Fig. 1. The function represents possible effects the two words have on each others sentiment by multiplying each one's matrix with the others vector.

$$\mathbf{p} = f(\mathbf{Ba}, \mathbf{Ab}) = g\left(\mathbf{W}\begin{bmatrix}\mathbf{Ba}\\ \mathbf{Ab}\end{bmatrix}\right), \quad \mathbf{W} \in \mathbb{R}^{n\times 2n}$$

We can then use the vector $\mathbf{p}$ to determine the sentiment of the phrase it represents. Instead of focusing on only two classes of sentiment (negative and positive), the model can predict a sentiment distribution over K classes. Applying the softmax function to $\mathbf{p}$ in combination with some weights $\mathbf{W}_{\text{class}}$, element-wise, gives us an estimate $\mathbf{d}$ of membership probability for each of the K sentiment classes:

$$\mathbf{d} = \operatorname{softmax}\left(\mathbf{W}_{\text{class}}\,\mathbf{p}\right), \quad \mathbf{W}_{\text{class}} \in \mathbb{R}^{K\times n}, \quad \mathbf{d} \in \mathbb{R}^{K}$$

$$\operatorname{softmax}_i(\mathbf{z}) = \frac{e^{\mathbf{z}_i}}{\sum_{i\neq j}^{n} e^{\mathbf{z}_j}}$$

To determine the amount of error between the reference and predicted sentiment probability distributions, $\mathbf{y} \in \boldsymbol{Y}$ and $\mathbf{d}$, respectively, we compute the binary cross entropy errors for each of the K classes. The loss function J is simply the mean error across all training instances:

$$E(\mathbf{y}, \mathbf{d}) = -\frac{1}{K}\sum_{i=1}^{K}\left(\mathbf{y_i}\ln\left(\mathbf{d_i}\right) + \left(1-\mathbf{y_i}\right)\ln\left(1-\mathbf{d_i}\right)\right)$$

$$J = \frac{1}{N}\sum_{i=1}^{N} E\left(\boldsymbol{Y}^{(i)}, \mathbf{d}^{(i)}\right)$$

While the initial vector components $\mathbf{x}$ of all the word MV-representations could be initialized to random values, we can also pretrain them, which has been shown to be beneficial for many tasks (Erhan et al., 2010). Following these insights, we initialize the vectors to word embeddings produced by *word2vec*,[1] an implementation of the skip-gram model by Mikolov et al. (2013), trained on the fHrWaC[2] corpus (Šnajder et al., 2013; Ljubešić and Erjavec, 2011).

[1] https://code.google.com/p/word2vec/
[2] http://takelab.fer.hr/data/fhrwac/

Similarly, we set all the initial word matrix components $\mathbf{X}$ to the identity matrix, adding a small amount of noise. Since $\mathbf{X} \approx \mathbf{I}$, it ceases to have an effect on the sentiment of a word when multiplied with that word's vector, as in the definition of function f. This ensures that words by default do not function as operators; they neither intensify, attenuate, nor flip the sentiment of the words they are paired with.

The model's total number of parameters equals $2n^2 + Kn + L(n+n^2)$, corresponding to sizes of $\mathbf{W}$, $\mathbf{W}_{\text{class}}$, and the MV-representations of all L words in the lexicon. We optimize these parameters by minimizing J with stochastic gradient descent, using a starting learning rate of $\alpha = 0.1$ and diminishing it linearly towards zero. Due to the large space complexity ($O(Ln^2)$), there are practical restrictions on the value of n. However, it has been shown that setting n to larger values (larger than 11) does not improve the performance (Socher et al., 2012).

4. Evaluation

We evaluate the model on two different datasets of phrases:[3] (1) a synthetic dataset where phrases have been assembled and their sentiment distributions labeled manually and (2) a dataset of manually translated common phrases extracted from movie reviews in English.

Since movies are commonly rated on a scale of 1 to 10, and indeed our source for the second dataset uses that very same rating scheme, we will be classifying phrases into K=10 sentiment classes that each correspond to a particular rating ranging from 1 (the worst) to 10 (the best). Additionally, we will use the same model trained for K=10 classes and apply it to classification of sentiment into K=3 classes.

4.1. Datasets

The datasets consist of unique two-word phrases paired with their sentiment distributions over a certain number K of classes. It should be noted that a reference sentiment distribution is never assigned to an individual word but exclusively to phrases. Each phrase occurs only once in a dataset, but an individual word may occur multiple times, as a part of different phrases (e.g., *"good"*).

Synthetic dataset. The first set consists of 1500 different phrases composed of Croatian words, assembled by pairing each of the 25 different adverbs with each of the 60 different adjectives. The set is divided into 1200 training phrases and 300 test phrases. Each of the phrases is manually labeled by a probability distribution over the K=10 sentiment classes, determined subjectively by a single author considering the phrase outside of context. None of the phrases have been labeled with ambiguous sentiment, meaning their sentiment probability distributions contain only one single maximum.

Movie reviews dataset. The second dataset is based on a publicly available dataset of bigrams extracted from movie reviews written in English.[4] Each of the phrases is associated with its frequency of occurrence within reviews with each of 10 different possible ratings. Note that here we

[3] Datasets are available from http://takelab.fer.hr/data/crophrasesent
[4] http://compprag.christopherpotts.net/iqap-experiments.html

assume a correlation between a review's rating and the sentiment of phrases expressed within it, and so use the frequencies of occurrence to construct for each unique phrase a probability distribution over K=10 sentiment classes. Such a simplistic assumption might not hold in all cases (e.g., a positive phrase might, for whatever reason, appear often in negatively scored reviews and vice versa). Each phrase that occurred in total at least 300 times was manually translated into Croatian by a single annotator using his subjective judgment. The translated phrases are then compiled into a dataset consisting of 1026 different phrases containing 208 unique words. The dataset is divided into a training set consisting of 821 and a test set consisting of 205 instances.

4.2. Results and discussion

We evaluate the MV-RNN model for several different sizes of the word vector (n = 8, 10, 13, and 15). We present the results using two different measures: (1) the F1-score and (2) the mean Kullback-Leibler divergence (KL-divergence). The KL-divergence measures the (dis)similarity between the reference and predicted probability distributions $\mathbf{y}$ and $\mathbf{d}$, respectively:

$$\mathrm{KL}(\mathbf{y}, \mathbf{d}) = \sum_i \mathbf{y}_i \ln \frac{\mathbf{y}_i}{\mathbf{d}_i}$$

We compute two F1-scores: (1) for K=10 classes and (2) for K=3 classes (the *positive*, *negative*, and *neutral* class). The F1-score for the K=3 case is derived from the results of the K=10 case, by splitting the sentiment probability distribution into three ranges ($1 \leq$ negative ≤ 3; $4 \leq$ neutral ≤ 7; $8 \leq$ positive ≤ 10), for which we sum the probabilities assigned to individual scores. Such binning allows us to evaluate the model in a commonly used *negative/neutral/positive* sentiment classification setting.

For the K=3 classification setting, we compare the MV-RNN against two baselines: a simple sentiment lexicon-based model (SentiLex) and a support vector machine (SVM) model. The SentiLex model assigns a positive ($+1$), negative (-1), or neutral (0) score to each word in a phrase, and then simply sums up these polarities. The SVM model is trained on a concatenation of two word vectors as features, either two one-hot vectors ($\mathrm{SVM}_{\text{1-hot}}$) or two 100-dimensional pretrained vectors ($\mathrm{SVM}_{\text{Pre}}$).

The evaluation results for the synthetic and movie review dataset are given in Tables 1 and 2, respectively. The MV-RNN models perform very well on the synthetic dataset, clearly outperforming the baselines. However, good performance on this dataset should come as no surprise, because the dataset is (1) very *clean* – there is no sentiment ambiguity (e.g., one phrase having high probabilities for both positive and negative scores) and (2) each word occurs in the dataset paired with every other and is found within different phrases many times. Individual words in real datasets will occur much less frequently. Reference and predicted probability distributions for four example phrases from the synthetic dataset are depicted in Fig. 2.

On the more realistic move reviews dataset, with significantly more sentiment ambiguity and a smaller number of occurrences of single words, the model performs worse than on the synthetic set. The performance is, nonetheless, well above the baselines for K=3, and comparable to the performance achieved by the same model for English (Socher et al., 2012). Example reference and predicted probability distributions are depicted in Fig. 3.

It is apparent from the results that the model can correctly capture the way words can intensify, attenuate, or flip entirely the sentiment inherent in words they are paired with. A lower performance on the movie reviews dataset may perhaps be traced down to the assumption upon which the reference distributions were created: phrases are negative if they more frequently occur in generally negative reviews and positive if they more frequently occur in positive reviews. However, an unambiguously negative phrase still may occur

		F1-score		
	n	K=3	K=10	KL
SentiLex	–	43.0	–	–
$\mathrm{SVM}_{\text{1-hot}}$	85	83.9	–	–
$\mathrm{SVM}_{\text{Pre}}$	100	91.8	–	–
$\mathrm{MV\text{-}RNN}_{\text{Rand}}$	8	93.0	63.1	0.025
$\mathrm{MV\text{-}RNN}_{\text{Rand}}$	10	90.1	71.6	0.025
$\mathrm{MV\text{-}RNN}_{\text{Rand}}$	13	92.7	70.0	**0.021**
$\mathrm{MV\text{-}RNN}_{\text{Rand}}$	15	**93.1**	69.6	**0.021**
$\mathrm{MV\text{-}RNN}_{\text{Pre}}$	8	91.2	68.7	0.026
$\mathrm{MV\text{-}RNN}_{\text{Pre}}$	10	92.8	**76.4**	0.023
$\mathrm{MV\text{-}RNN}_{\text{Pre}}$	13	91.2	74.6	0.024
$\mathrm{MV\text{-}RNN}_{\text{Pre}}$	15	92.4	74.8	0.023

Table 1: Results for the **synthetic** dataset, using random (Rand) and pretrained (Pre) initial vectors.

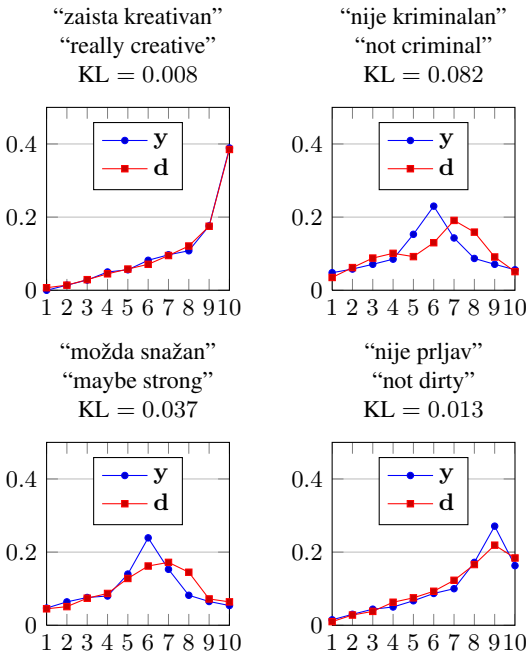


Figure 2: Results for selected phrases from the **synthetic** test set. The x-axis shows the K=10 sentiment classes, while the y-axis shows the sentiment probability distribution (the probability of the phrase belonging to a specific sentiment class). Reference sentiment probability distributions are shown in blue and classifier predictions in red.

		F1-score		
	n	K=3	K=10	KL
SentiLex	–	45.2	–	–
$SVM_{1\text{-hot}}$	134	63.8	–	–
SVM_{Pre}	100	61.2	–	–
$MV\text{-}RNN_{Rand}$	8	68.9	34.6	0.055
$MV\text{-}RNN_{Rand}$	10	67.2	36.5	0.055
$MV\text{-}RNN_{Rand}$	13	**69.2**	34.9	0.056
$MV\text{-}RNN_{Rand}$	15	67.8	40.8	**0.054**
$MV\text{-}RNN_{Pre}$	8	63.7	33.3	0.065
$MV\text{-}RNN_{Pre}$	10	67.6	38.3	0.066
$MV\text{-}RNN_{Pre}$	13	64.3	**43.1**	0.067
$MV\text{-}RNN_{Pre}$	15	67.7	37.1	0.066

Table 2: Results for the **movie reviews** dataset, using random (Rand) and pretrained (Pre) initial vectors.

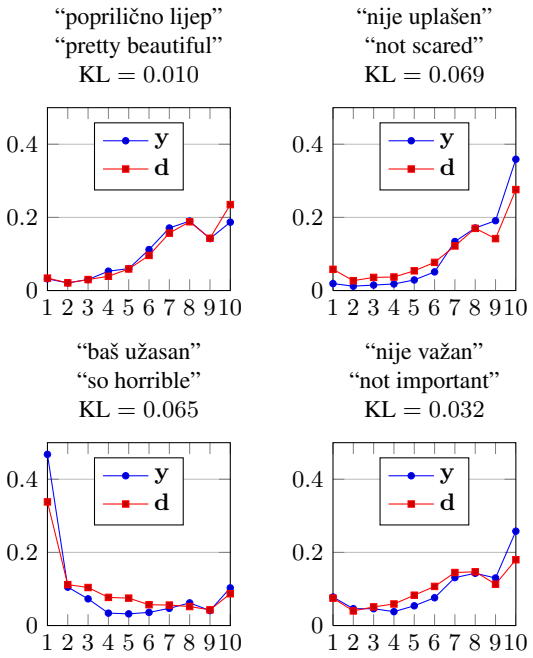


Figure 3: Results similar to those from Fig. 2, but for chosen phrases from the **movie reviews** test set.

in an otherwise very positive review with a high rating, and vice-versa. Similarly, a phrase may be ambiguous in that it can be used in both positive and negative contexts. These ambiguities are likely to affect the model's performance.

Surprisingly, pretraining the word vectors does not improve the performance. Moreover, in some cases having word vectors pretrained actually degrades performance. This is likely due to the fact that pretraining serves to learn the semantic meaning of the words, which may often conflict with their sentiment. For example, two antonyms will, after pretraining, have similar word vector representations, but their sentiment is directly opposite (e.g., *"better"* vs. *"worse"*).

5. Conclusion

While lexicons of prior sentiment are useful in many sentiment analysis tasks, multiword phrases often have a sentiment different from the prior sentiment of their constituent words. In this paper we used a deep neural network model proposed by Socher et al. (2012) to learn the sentiment of two-word Croatian phrases. We evaluated the model on two different datasets: one synthetic and the other realistic. Experimental results suggest that deep learning models are well-suited for the task of modeling the sentiment of Croatian phrases, confirming previous results for English.

We have not exploited the key capability of the MV-RNN model: the recursive application to arbitrary length n-grams, which has been shown to be very effective for modeling the sentiment of complete sentences (Socher et al., 2013). We intend to pursue this line of work and experiment with predicting the sentiment of complete sentences in Croatian.

6. References

M. Baroni and R. Zamparelli. 2010. Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space. In *Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193. ACL.

D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio. 2010. Why does unsupervised pretraining help deep learning? *Journal of Machine Learning Research*, 11:625–660.

N. Ljubešić and T. Erjavec. 2011. hrWaC and slWac: compiling web corpora for Croatian and Slovene. In *In Proc. of Text, Speech and Dialogue 2011*, Lecture Notes in Computer Science, pages 395–402. Springer.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

J. Mitchell and M. Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

J. Šnajder, S. Padó, and Ž. Agić. 2013. Building and evaluating a distributional memory for Croatian. In *51st Annual Meeting of the Association for Computational Linguistics*, pages 784–789. ACL.

R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 151–161. ACL.

R. Socher, B. Huval, C. D. Manning, and A. Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proc. of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211. ACL.

R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642. ACL.

A. Yessenalina and C. Cardie. 2011. Compositional matrix-space models for sentiment analysis. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 172–182. ACL.

HEIDELTIME.HR: Extracting and Normalizing Temporal Expressions in Croatian

Luka Skukan, Goran Glavaš, Jan Šnajder

University of Zagreb, Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
{luka.skukan, goran.glavas, jan.snajder}@fer.hr

Abstract

Temporal expression extraction and normalization are important for many NLP tasks and have been the topic of extensive research. While the majority of research on temporal expression extraction was performed for English, there has recently also been work on temporal processing for other languages. In this paper, we describe HEIDELTIME.HR, the Croatian resources for HeidelTime – a multilingual, cross-domain temporal expression tagger. HeidelTime recognizes temporal expressions in text and normalizes them according to the TIMEX3 annotation standard. We compile WikiWarsHr, a corpus of historical narratives in Croatian manually annotated for temporal expressions. On WikiWarsHr, HEIDELTIME.HR achieves results comparable to those originally achieved by HeidelTime on English texts, with F1-scores of 0.93 and 0.86 for expression extraction and normalization, respectively.

HEIDELTIME.HR: luščenje in normaliziranje časovnih izrazov v hrvaščini

Luščenje in normalizacija časovnih izrazov sta pomembna za raznovrstne naloge s področja računalniške obravnave naravnega jezika in sta bila predmet številnih raziskav. Medtem ko je bila večina raziskav luščenja časovnih izrazov opravljenih za angleščino, pa so bile v zadnjem času raziskave izvedene tudi za druge jezike. V prispevku opišemo HeidelTime.Hr, hrvaške vire za HeidelTime – večjezični in prekdomenski označevalec za časovne izraze. HeidelTime prepozna časovne izraze v besedilu in jih normalizira glede na standard za označevanje TIMEX3. Izdelamo WikiWarsHr, korpus zgodovinskih pripovedi v hrvaščini, ki je bil ročno označen za časovne izraze. Na WikiWarsHr doseže HeidelTime.Hr rezultate, primerljive s tistimi, ki jih je HeidelTime dosegal na angleških besedilih, z mero F 0,93 za luščenje in 0,86 za normalizacijo časovnih izrazov.

1. Introduction

The ability to extract and normalize temporal expressions in natural language texts is of major importance for natural language processing tasks, such as summarization and question answering, but also for reasoning about events and time in general. Temporal expression extraction is the task of identifying temporal expressions and their extent. The normalization task amounts to turning extracted temporal expressions into a fully specified value and formatting them according to some standard, including under-specified values.

While a number of temporal taggers are available, mostly for English and other major languages, a temporal expression tagger for Croatian does not yet exist. A new temporal expression tagger could be implemented, or an existing multilingual system could be adapted to work for Croatian. We chose the latter approach in this work, building on an existing and widely used framework.

In this paper, we describe HEIDELTIME.HR, the Croatian resources for the rule-based temporal expression tagger HeidelTime (Strötgen et al., 2013).[1] HeidelTime extracts and normalizes temporal expressions according to the TIMEX3 standard (Pustejovsky et al., 2003), and emerged as a winner in the TempEval-2 (Verhagen et al., 2010) and TempEval-3 (UzZaman et al., 2012) shared evaluation tasks. HeidelTime is a multilingual tagger, with resources been developed for English, German (Strötgen et al., 2013), Arabic, Italian, Spanish, Vietnamese (Strötgen et al., 2014a), French (Moriceau and Tannier, 2014), Chinese (Li et al., 2014), Dutch, and Russian. We have developed Croatian resources, which will be included in the next HeidelTime release.[2]

To develop and evaluate the tagger, we compiled WikiWarsHr, a corpus of historical narratives in Croatian manually annotated for temporal expressions. On this corpus, HEIDELTIME.HR achieves results comparable to those originally achieved by HeidelTime on English texts.

The structure of this paper is as follows. We describe the mechanisms of HeidelTime in Section 2. Section 3 describes the HEIDELTIME.HR resources. In Section 4, we describe the WikiWarsHr corpus and present the evaluation results. Section 5 concludes the paper.

2. HeidelTime tagger

The HeidelTime tagger extracts and normalizes temporal expressions according to the TIMEX3 standard (Pustejovsky et al., 2003). In TIMEX3, each temporal expressions is assigned a Type and a Value. A Type may be a *Date*, *Time*, *Duration* or *Set*. The Value corresponds to a temporal value, partially dependent on Type (e.g. a Date "*2014-10*" for October of 2014).

HeidelTime features a generic, language-independent core, written in Java, and a language-dependent part, the so-called language resources. A language resource consist of three sets: (1) expression resources , (2) normaliza-

[1] http://code.google.com/p/heideltime

[2] The HEIDELTIME.HR resources are also available from http://takelab.fer.hr/heideltimehr

(DCT: June 21st 2014)

The field of AI research was founded at a conference on the campus of Dartmouth College in the <TIMEX3 tid="t1" type="DATE" value="1956-SU">summer of 1956</TIMEX3>. <TIMEX3 tid="t2" type="DATE" value="2014">58 years later</TIMEX3>., we still haven't achieved many of the goals proposed there. Still, artificial intelligence has advanced and is <TIMEX3 tid="t3" type="DATE" value="2014-06-21">today</TIMEX3> a part of our daily lives without most of us knowing it.

Figure 1: Example of under-specification resolution.

tion resources, and (3) rule resources. Expression resources are regular expressions used for extraction temporal expressions from text, e.g., phrases for months, weekdays, numbers, etc. Normalization resources translate matched tokens to their canonical form, according to TIMEX3, by applying normalization mapping to extracted patterns (e.g., "*May*" → "*05*"). Finally, the rule resources combine the previous two resources to extract and normalize temporal expressions. These may be complemented with additional regular expressions to form more complex match-and-normalize rules, e.g., for discarding parts of extracted expressions or for adding a modifier ("*early*", "*middle*", etc.).

Normalization is performed both on fully specified expressions ("*June 28, 1995*") and relative temporal expressions ("*tomorrow*"). The latter are expressions that cannot be normalized without contextual information. Normalization of relative temporal expressions is performed by leaving the expressions under-specified and relying on HeidelTime's generic focus-tracking system to assign them a more specific value. For example, given a document creation time (DCT) of June 20th, 2014, the expression "*tomorrow*" might be resolved as "2014-06-21". This step is performed by taking into account the type of the document (narrative, news, scientific, or colloquial) and the tenses of the verbs used in the sentence containing the under-specified temporal expression. Either the DCT or a previously mentioned value can be used in under-specified expression normalization, depending on the document type and the normalization rule. An example of resolving under-specified dates using both DCT and current focus is shown in Fig. 1. Additionally, HeidelTime supports functionality extensions in form of text post-processors written as Java code. These allow for more verbose expression resolution, e.g., computing the date of lunar holidays such as Easter.

3. HeidelTime.hr

The task of developing resources for Croatian language consisted of developing three above-mentioned sets of resources. We next describe the resources and the development methodology.

3.1. Preprocessing

HeidelTime requires text to be pre-annotated with token, sentence and part-of-speech (POS) information. We used the CSTLemma lemmatiser (Jongejan and Haltrup, 2005) for token splitting and lemmatization,[3] and the HunPos part-of-speech tagger (Halácsy et al., 2007) to obtain the POS information. To integrate this functionality with HeidelTime, we wrote a Java wrapper that allows the tagger's engine to invoke it during pre-processing. HunPos and CSTLemma were previously trained to work with Croatian texts (Agić et al., 2013).

3.2. Resources

HeidelTime.hr resources are divided into several classes. The expression and normalization resources are divided into descriptive classes, according to their common roles in temporal constructs, with each normalization resource corresponding to an expression resource. Some examples include the *MonthLong* and *Timezone* resources. The rules are divided according to their semantics in the TIMEX3 standard into *Date*, *Time*, *Duration* and *Set* resources. Altogether, there are 199 rule resources for Croatian: 123 for dates, 37 for time, 24 for durations, and 15 for sets. This number is much larger than for English, but of comparable size to resources for other inflected languages, such as French, which has 157 rule resources (Moriceau and Tannier, 2014). Furthermore, as a highly inflected language, Croatian requires a large number of rule variations to account for the inflections. This issue could have been partially avoided by using lemmas instead of raw words. However, we chose not to do so for three reasons: (1) The implementation would be complex and time-consuming; (2) Due to the generic nature of the HeidelTime engine, lemmatization would have to be integrated system-wide, and the decision of whether to use lemmatization would have to be specified for each set of language-specific resources; (3) Errors in lemmatization would propagate into HeidelTime, decreasing its accuracy.

As an illustration, consider the following example of a complete HeidelTime extraction rule, which can be used to extract and normalize parts of seasons, such as "*ranog proljeća* ("*early spring*"):

```
RULENAME="date_r9b",
EXTRACTION="%rePartWords_g1 %reSeason_g2",
NORM_VALUE="UNDEF-year-%normSeason(g2)",
NORM_MOD="%normPartWords(g1)"
```

The extraction part of the rule extracts expressions describing a specific part of something (e.g., "*early*", "*middle of*", etc.) and stores it as *group 1* (g_1), as well as an expression denoting a season (e.g., "*summer*"), which is stored as *group 2* (g_2). It leaves the year undefined as "UNDEF-year", which will be resolved by HeidelTime using the temporal context of the sentence. The "part word" in g_1 is normalized as part of the modifier (NORM_MOD), which makes the value more specific. The season, g_2, is combined with the determined year to get the temporal value of the expression. Assuming the inferred year is 2014, the given expression "*ranog proljeća*" would

[3]The lemmas produced by the CSTLemma lemmatiser are presently not used by the system, but may be integrated in the future (cf. Section 3.2.).

be normalized as *<TIMEX3 tid="t1" value="2014-SP" mod="START">ranog proljeća</TIMEX3>*.

3.3. Development methodology

We developed HEIDELTIME.HR in two phases. We first translated the existing English and German resources (Strötgen et al., 2013) into Croatian, wherever appropriate. We then used a data-driven approach to further develope and refine the resources, using a subset of manually-annotated Wikipedia corpus (cf. Section 4.1.) as a development set. The development set consists of ten Wikipedia articles of varying length, altogether containing 29,563 non-punctuation tokens and 677 temporal expressions. Usage examples for all TIMEX3 types of rule resources are given in Fig. 2.

(a) Službeno, američki angažman je završio u <TIMEX3 tid="t128" type="TIME" value="2010-08-31T17:00">utorak, 31. kolovoza, u 17:00 sati</TIMEX3>. Otprilike 50.000 vojnika je ostalo u Iraku do <TIMEX3 tid="t129" type="DATE" value="2011" mod="END">kraja 2011.</TIMEX3>

(Officially, the American engagement ended on Tuesday, the 31st of October, at 5:00 PM. Around 50,000 soldiers stayed in Iraq until the end of 2011.)

(b) Rat nije bitnije promijenio granicu između dvije države, no cijena <TIMEX3 tid="t23" type="DURATION" value="P8Y">osmogodišnjeg</TIMEX3> ratovanja u ljudskim žrtvama i posljedicama po gospodarstvo je bila ogromna i za Irak i za Iran.

(The war did not result in major changes in the border between the two states, but the price of an eight-year war, in human lives and damage to the economy, was great for both Iraq and Iran.)

(c) Proizvodnja žita je opadala prosječno 3,5% <TIMEX3 tid="t55" type="SET" value="P1Y">godišnje</TIMEX3> između <TIMEX3 tid="t53" type="DATE" value="1978">1978.</TIMEX3> i <TIMEX3 tid="t54" type="DATE" value="1990">1990.</TIMEX3> zbog borbi, nestabilnosti u seoskim područjima, duge suše i propale infrastrukture.

(The wheat production dropped, on average, 3.5% a year between 1978 and 1990, due to fighting, instability in rural areas, the long drought and the ruined infrastructure.)

Figure 2: Examples of Croatian documents tagged with HeidelTime.

When develeoping the rules, there were a few corner cases that we deliberately chose to ignore. More specifically, to not warrant a rule, an expression had to satisfy one of the following conditions:

1. A rule that would have been written to match the expression would be imprecise (i.e., result in more false positives than true positives). E.g., "*skoro* ("*soon/almost*") is more often an adverb of degree than a temporal expression;

2. An expression is complex or unique, and therefore unlikely to appear in other documents, such as "*nedjelju oko 14,45 sati po srednjoeuropskome vremenu* ("*Sunday at about 2,45 PM according to Central European time*");

3. An expression is very domain-specific and would potentially lead to a performance decrease across the board (e.g., references to the beginning or end of a particular war).

The rationale for the first condition is straightforward: recall would rise, but precision would plummet. The second and third conditions are meant to prevent overfitting. While including the specific rules would slightly increase the performance on the development set, it would not increase or could potentially decrease the performance on unseen data.

4. Evaluation

As part of this work, we have compiled WikiWarsHR, a corpus of Croatian Wikipedia manually annotated for temporal expressions. We used this corpus for the development and evaluation of HEIDELTIME.HR.

4.1. WikiWarsHR corpus

WikiWarsHR is inspired by the WikiWars corpus of Mazur and Dale (2010). While the content is similar (21 out of 22 articles detail the same wars as the original WikiWars corpus), the difference is that we chose to annotate WikiWarsHr using TIMEX3, a subset of the TimeML standard (Pustejovsky et al., 2003). The entire corpus contains almost 60,000 non-punctuation tokens and 1,440 temporal expressions in 22 articles. Two of these articles mostly contained historic (BC) temporal expressions, the processing of which is the newest addition to HeidelTime (Strötgen et al., 2014b) and which we have not yet implemented for Croatian. Therefore, we excluded these two articles, and divided the remaining 20 articles into the development and test set. This gave us a test set consisting of 10 articles, containing 21,644 tokens and 609 tagged temporal expressions. The articles are very diverse in length and temporal expression density, ranging from the minimum of 235 tokens and 12 tagged temporal expressions up to 9,722 tokens and 181 tagged temporal expressions. The WikiWarsHr corpus is freely available.[4]

4.2. Experimental setup

We computed the precision, recall, and F1-score for both expression extraction and normalization. The scores were computed on the expression level. We used two evaluation settings: *relaxed* and *strict*. In the relaxed setting,

[4]Available under the Creative Commons BY-NC-SA license from `http://takelab.fer.hr/wikiwarshr`

	Extraction			Type			Value		
Dataset & Tagger	P	R	F1	P	R	F1	P	R	F1
CroNER	0.83	0.54	0.66	0.82	0.54	0.65	–	–	–
HeidelTime (Development set)	0.95	0.97	0.96	0.94	0.96	0.95	0.86	0.88	0.87
HeidelTime (Test set)	0.94	0.96	0.95	0.93	0.95	0.94	0.86	0.88	0.87

Table 1: Tagger performance on WikiWarsHw corpus (relaxed match).

	Extraction			Type			Value		
Dataset & Tagger	P	R	F1	P	R	F1	P	R	F1
CroNER	0.26	0.17	0.21	0.26	0.17	0.21	–	–	–
HeidelTime (Development set)	0.93	0.95	0.94	0.93	0.95	0.94	0.86	0.88	0.87
HeidelTime (Test set)	0.92	0.93	0.93	0.91	0.93	0.92	0.85	0.87	0.86

Table 2: Tagger performance on WikiWarsHr corpus (strict match).

even partial matches are considered to be valid extraction and their normalization is scored. A partial match pertains to cases in which the tagged expression and the gold standard share at least one token. In the strict setting, only complete matches are considered correct. We computed the normalization scores for both the *Type* and *Value* properties of temporal expressions.

As a baseline, we evaluate CroNER (Glavaš et al., 2012) – a named entity recognition and classification system for Croatian – on the WikiWarsHr corpus. CroNER is capable of identifying temporal expressions belonging to *Date* and *Time* TIMEX3 types. As CroNER cannot normalize temporal expressions, we only evaluated expression extraction and type normalization. We measured the CroNER's performance on the entire WikiWars corpus (a union of the development and test set).

4.3. Results

Table 1 gives evaluation results for relaxed match. Extraction and normalization scores are high, particularly for the *Type*, with a negligible performance drop on the test set. Table 2 shows strict evaluation results for the two sets. The differences in the results compared to relaxed evaluation are almost negligible, with the drop in performance of 2% or less. This indicates that most errors are caused by errors in value normalization, rather than expression extraction.

Overall, the results are quite satisfying and comparable to those achieved by HeidelTime for English (Extraction 0.9; Type 0.82; Value 0.78) and Spanish (Extraction 0.9; Type 0.87; Value 0.85) (Strötgen et al., 2013).[5] However, part of this success can probably be attributed to the simpler nature of WikiWarsHr corpus in comparison to its English counterpart and a relatively large number of rules, many written primarily for the historical narratives.

4.4. Error analysis

As discussed above, most errors stem from value normalization. The few extraction errors are usually caused by unrecognized references to events or unique, large expressions. This is mostly due to the nature of the narratives – times relative to referenced events, implicitly switching focus between years, etc. Examples of both types of errors are given in Fig. 3.

Extraction error:

. . . nakon japanskog napada na Pearl Harbor . . .

(After the Japanese attack on Pearl Habor)

Normalization error:

Veljača, ožujak i travanj su bili relativno mirni mjeseci u usporedbi s krvavim studenim i siječnjom. . .

(The months February, March and April were relatively calm compared to the bloody November and January. . .)

Figure 3: Examples of tagging errors (expressions on which the errors occur are underlined).

Type	Errors	Occurrences	Error (%)
Date	85	1132	7.5
Time	1	23	4.3
Duration	1	50	2
Set	3	5	60

Table 3: Value normalization errors according to type.

Due to majority of errors originating from value normalization, we made a breakdown of normalization errors by expression type, on the union of the testing and development datasets. We considered only the expressions that have been correctly extracted (using strict evaluation) and had their type correctly normalized to match their counterparts in the gold standard. Table 3 shows that the largest number of errors stem from *Date* values, which also ac-

[5]Here we refer to the best results achieved on TempEval-3 datasets, obtained using tuned rulesets and relaxed matching.

count for the majority of temporal expressions in the corpus. Most of these errors are due to using a wrong focus point during normalization. Normalization of *Time* and *Duration* expression performs better, with a lower than 5% error rate. Value normalization performs poorly for *Set* values, with three out of five values normalized incorrectly. However, all three errors can be traced down to a systematic inconsistency: HeidelTime tagged all occurrences of "yearly" with the value "XXXX" (denoting "*every* year"), whereas the human annotators tagged it as "P1Y" (denoting "*once per* year"). In this case, however, the two tags are semantically equivalent.

5. Conclusion

We presented HEIDELTIME.HR, a resource we developed for temporal tagging of Croatian texts with the multilingual temporal tagger HeidelTime. We also described WikiWarsHr, a new Wikipedia-based corpus of Croatian historical narratives manually annotated for temporal expressions. On WikiWarsHr corpus, HEIDELTIME.HR achieves an F1-score of 0.93 and 0.86 for temporal expression extraction and normalization, respectively. This result is comparable to the result of HeidelTime for English.

Future extensions of the presented HiedelTime resources will include incorporating rules for historic dates, the newest addition to HeidelTime, into HEIDELTIME.HR. Furthermore, potential improvements in performance and ease of use could be achieved by adapting HeidelTime to work with lemmas instead of wordforms.

6. References

Ž. Agić, N. Ljubešić, and D. Merkler. 2013. Lemmatization and Morphosyntactic Tagging of Croatian and Serbian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing (BSNLP 2013)*. Association for Computational Linguistics.

G. Glavaš, M. Karan, F. Šarić, J. Šnajder, J. Mijić, A. Šilić, and B. Dalbelo Bašić. 2012. CroNER: A State-of-the-Art Named Entity Recognition and Classification for Croatian. *Information Society*.

P. Halácsy, A. Kornai, and C. Oravecz. 2007. Hunpos: an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 209–212. Association for Computational Linguistics.

B. Jongejan and D. Haltrup. 2005. The CST Lemmatiser. *Center for Sprogteknologi, University of Copenhagen version*, 2.

H. Li, J. Strötgen, J. Zell, and M. Gertz. 2014. Chinese Temporal Tagging with HeidelTime. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 133–137.

P. Mazur and R. Dale. 2010. WikiWars: A new corpus for research on temporal expressions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 913–922. Association for Computational Linguistics.

V. Moriceau and X. Tannier. 2014. French resources for extraction and normalization of temporal expressions with HeidelTime. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA).

J. Pustejovsky, J. M. Castano, R. Ingria, R. Sauri, R. J. Gaizauskas, A. Setzer, G. Katz, and D. R. Radev. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. *New directions in question answering*, 3:28–34.

J. Strötgen, J. Zell, and M. Gertz. 2013. HeidelTime: Tuning English and Developing Spanish Resources for TempEval-3. In *Proc. of Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 15–19.

J. Strötgen, A. Armiti, T. Van Canh, J. Zell, and M. Gertz. 2014a. Time for More Languages: Temporal Tagging of Arabic, Italian, Spanish, and Vietnamese. *ACM Transactions on Asian Language Information Processing (TALIP)*, 13(1):1.

J. Strötgen, T. Bögel, J. Zell, A. Armiti, T. V. Canh, and M. Gertz. 2014b. Extending heideltime for temporal expressions referring to historic dates. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2390–2397. European Language Resources Association (ELRA).

N. UzZaman, H. Llorens, J. Allen, L. Derczynski, M. Verhagen, and J. Pustejovsky. 2012. Tempeval-3: Evaluating events, time expressions, and temporal relations. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantcis (*SEM 2013)*. Association for Computational Linguistics.

M. Verhagen, R. Sauri, T. Caselli, and J. Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62. Association for Computational Linguistics.

Merjenje berljivosti strojnih prevodov s sledilcem očesnih gibov

Kristijan Armeni,* Grega Repovš,† Špela Vintar‡

* Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
kristijan.armeni@gmail.com
† Oddelek za psihologijo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
grega.repovs@psy.ff.uni-lj.si
‡ Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
spela.vintar@ff.uni-lj.si

Povzetek

V prispevku predstavimo študijo, s katero smo želeli preveriti uporabnost sledilca očesnih gibov kot dopolnilno metodo za evalvacijo strojnih prevodov iz angleščine v slovenščino. Najprej predstavimo priložnostno kategorizacijo napak v strojnih prevodih, na podlagi katere smo pripravili tri vrste prevodov (popravljene, delno popravljene in nepopravljene). Med tem ko so udeleženci brali besedila za razumevanje, smo posneli gibanje oči. Po branju je vsak udeleženec podal še subjektivne ocene prebranega besedila. Udeleženci so sicer vse tri tipe prevodov ocenili kot kvalitativno različne, rezultati sledilca pa pokažejo razlike le med branjem povsem popravljenih in nepopravljenih prevodov, ne pokažejo pa razlik med branjem delno popravljenih in nepopravljenih prevodov.

Abstract: Eye tracking measures as indicators of readability for machine-translated texts

In the present paper, we present a study where we tested the newly proposed methodology for evaluating machine-translated texts (English-Slovene language pair). We first outline our ad hoc error categorization scheme, which served as the basis for the preparation of three types of translations (raw MT output, partially corrected, and completely corrected). Participants were instructed to read the texts for comprehension while their eye movements were recorded. After reading, they were asked to give their subjective evaluations as well. While the three types of translations were given significantly different scores by the participants, eye-tracking measures show difference between completely corrected and non-corrected MT output, but no difference between partially and non-corrected MT output.

1. Uvod

Namen študije, predstavljene v tem prispevku, je bil test sledilca očesnih gibov kot metodologije za evalvacijo strojnih prevodov (jezikovni par angleščina-slovenščina).[1] Sledilec očesnih gibov je v psiholingvistiki standardna metodologija za raziskovanje kognitivnih procesov med branjem, z razvojem tehnologije pa se uporaba prenaša na sorodna področja. Množica empiričnih dognanj raziskav tekom zadnjih nekaj desetletij kaže, da je gibanje oči med branjem tesno povezano s sočasnimi kognitivnimi procesi pri bralcih (ang. *online cognitive processes*, Rayner, 1998; Just et al., 1980).

Osnovna hipoteza te študije je, da predstavljajo manj berljiva besedila, ki vsebujejo veliko napak (npr. zelo slab strojni prevod), za bralca večje kognitivno breme kot bolj berljiva besedila brez napak. Če obstaja povezava med kognitivnimi procesi ter gibanjem oči med branjem, potem bi morale specifične mere sledilca očesnih gibov (npr. število fiksacij, čas branja, povprečno trajanje fiksacije) odražati razlike v berljivosti besedil. Ker na ta način posnamemo branje strojnih prevodov v realnem času, bi takšne mere lahko predstavljale zanimivo, dodatno informacijo o kvaliteti besedil.

2. Evalvacija strojnih prevodov

Na področju strojnih prevajalnikov je evalvacija sestavni del razvojnega cikla posameznega sistema. Obstaja več metod evalvacije strojnih prevodov, pri čemer ima vsaka svoje prednosti in pomanjkljivosti. Metode za evalvacijo strojnih prevodov se v grobem delijo v dve skupini: ročne in avtomatske.

Metoda ročne evalvacije, tj. mnenje človeških ocenjevalcev oz. končnih uporabnikov, je tista, ki velja za referenčno in ki najbolje odseva kvaliteto strojnega prevoda. Zelo pogosto omenjeni pomanjkljivosti ročnih evalvacij sta cena ter zamudnost (Papineni et al., 2002; Callison-Burch et al., 2007; Koehn, 2010; Verdonik in Sepesy Maučec, 2013); obsežne evalvacijske delavnice z veliko udeleženci namreč predstavljajo nezanemarljive stroške. V zadnjem času se zato evalvacije opravlja tudi z množičnim zbiranjem podatkov preko spletnih platform (Graham et al., 2013b).

Druga velika težava človeških ocen je neskladnost oziroma t. i. stopnja (ne)strinjanja med posameznimi ocenjevalci (ang. *inter-annotator agreement*). Do neskladnosti pride tudi pri vsakem posamezniku, zato se dodatno preverja še ocenjevalčevo konsistentnost (ang. *intra-annotator agreement*). Pri ocenjevanju se najpogosteje uporabljajo pet in sedem stopenjske lestvice (Callison-Burch et al., 2007), pri čemer se v praksi izkaže, da imamo ljudje zelo različne kriterije in smo lahko različno strogi. Pojavi se torej dilema, ali ne povedo takšni rezultati več o samih ocenjevalcih kot pa o besedilih. Vrednost tako dobljenih rezultatov je lahko vprašljiva, posledično pa tudi sami zaključki, do katerih naj bi vodili (Koehn, 2010; Graham et al. 2013a).

Glavni prednosti avtomatske metode sta hitrost in cenovna dostopnost, vendar pa se avtomatske metrike se še vedno smatrajo le za nepopolno nadomestilo človeške

[1] Vsebina tega prispevka povzema ideje in rezultate, predstavljene v magistrski nalogi (2014) z istim naslovom.

evalvacije in bodo namreč označene za učinkovite le, če bodo podale rezultate, ki visoko korelirajo z rezultati človeške evalvacije (Callison-Burch et al., 2007; Verdonik in Sepesy Maučec, 2013; Graham et al., 2013b).

Avtomatske metrike imajo še druge pomanjkljivosti, med drugim predvsem to, da je končni rezultat takšne evalvacije zgolj specifična številčna vrednost, ki odseva več različnih parametrov. Ker je v postopke računanja končne vrednosti vključenih več faktorjev, pomen posamezne številke ni jasen (Koehn, 2010). Poleg tega je pri takšnih avtomatskih metrikah zelo pomembno, kakšen je referenčni prevod (ali več prevodov), ki ga vključimo v evalvacijsko gradivo (Verdonik in Sepesy Maučec, 2013).

V znanstveni literaturi o evalvaciji tako novosti in predlogi za izboljšave obstoječih tehnik niso redkost. Nedavno je v raziskovalni skupnosti prišlo do ideje, da bi kot vir podatkov o berljivosti strojnih prevodov in tako posredno tudi kot orodje za evalvacijo lahko uporabili sledilec očesnih gibov.

3. Predhodne študije

Za izhodišče smo vzeli dve nedavni študiji, kjer so pokazali, da bi sledilec očesnih gibov lahko uporabili kot dopolnilo standardnim tehnikam za evalvacijo strojnih prevodov, kot so npr. avtomatske metrike in človeške ocene. Dohertyja in soavtorje (2010) je zanimalo osnovno vprašanje, ali bo na podlagi mer sledilca očesnih gibov možno razlikovati med slabimi strojnimi prevodi in dobrimi strojnimi prevodi. Udeleženci v eksperimentu so brali predhodno ocenjene strojno prevedene stavke, ki so bili označeni kot »dobri« oziroma »slabi«. Rezultati so pokazali, da je bilo število fiksacij višje in povprečen čas branja daljši za slabo ocenjene kot za dobro ocenjene strojne prevode, medtem ko pri povprečnem trajanju fiksacije in razširjenju zenice ni prišlo do razlik. Kvaliteta posameznih stavkov je bila zmerno negativno korelirana z rezultati sledilca očesnih gibov.

Stymne in soavtorji (2012) so idejo nadgradili in pripravili nekoliko drugačen eksperiment. Zanimalo jih je, ali bi lahko sledilec očesnih gibov uporabili za analizo napak v strojnih prevodih. Udeleženci so brali cela besedila, ki so bila strojno prevedena s tremi različnimi statističnimi prevajalniki. V vsakem besedilu so označili dele z napakami in dele besedila, kjer napak ni bilo. Rezultati so pokazali razlike v številu fiksacij in povprečnem času pogleda (ang. *average gaze time*) za dele besedila, kjer so bile napake in kjer jih ni bilo. Do sistematičnih razlik (povprečen čas pogleda) je prišlo tudi med napakami pri posameznih strojnih prevajalnikih, prav tako tudi pri posameznih tipih napak, pri čemer je branje najbolj otežil napačen vrstni red besed.

Naša študija gradi na obeh opisanih raziskavah, a se od obeh tudi razlikuje. Podobno kot je pojasnjeno v obeh študijah, tudi našo v osnovi razumemo kot metodološko. V prvi študiji (Doherty et al., 2010) so udeleženci brali strojno prevedene stavke v dveh skupinah, medtem ko smo v našem primeru uporabili cela besedila in dodaten tip besedila med »dobrim« in »slabim«. S tega vidika je naša raziskava bolj primerljiva z drugo študijo (Stymne et al., 2012), kjer so uporabili cela besedila in primerjali napake v prevodih treh strojnih prevajalnikov. Po branju so podobno zbrali ocene udeležencev, a niso uporabili kontinuiranih lestvic kot v našem primeru. Rezultate sledilca so nato primerjali z rezultati avtomatskih metrik, česar sami nismo vključili v zasnovo.

4. Kategorizacija napak

V prvi fazi smo opravili priložnostno kategorizacijo napak v strojnih prevodih za jezikovni par angleščina-slovenščina. V naš profil napak smo vključili naslednje kategorije: ujemalne napake, pomenske napake, stilistične napake, pravopisne napake, napačen besedni red, izpuščene in vrinjene besede ter neprevedene besede.

Največ primerov napak smo uvrstili v kategorijo t.i. ujemalnih napak, kar je do neke mere pričakovano, saj je slovenščina v primerjavi z angleščino morfološko precej bolj razčlenjena. Pri jezikovnih parih, kjer je eden od jezikov morfološko bogat, pride v primeru statističnih prevajalnikov do »problema redkih podatkov« (ang. *sparse data problem*, Koehn 2010). Tabela 1 povzema pojavnost posameznih tipov napak za tri besedila iz preizkušnje.

	besedilo 1		besedilo 2		besedilo 3	
ujemalne napake	18	45 %	22	51 %	21	40 %
pomen	3	7,5 %	4	9,3 %	5	9,6 %
izpust	4	10 %	3	7 %	1	1,9 %
vrinjena beseda	1	2,5 %	2	4,7 %	5	9,6 %
besedni red	4	10 %	2	4,7 %	7	13,5 %
stilistika	5	12,5 %	6	14 %	8	15,4 %
neprevedeno	0	0 %	0	0 %	0	0 %
pravopis	5	12,5 %	4	9,3 %	5	9,6 %
skupaj	**40**	**100 %**	**43**	**100 %**	**52**	**100 %**

Tabela 1: Pojavnost posameznih tipov napak

Za našo študijo je kategorija ujemalnih napak ključnega pomena. Če v besedilih popravimo samo ujemalne napake, popravimo glede na vrednosti v zgornji tabeli v povprečju približno polovico vseh napak. Predstavljene številke velja interpretirati s previdnostjo; zaradi majhnega vzorca, specifičnega tipa besedil in enega samega ocenjevalca na podlagi teh podatkov ni moč posploševati o kvaliteti prevajalnika. Menimo pa, da je kategorija ujemalnih napak ustrezen kriterij za sistematično pripravo besedil. Besedilo brez ujemalnih napak bi se moralo po berljivosti uvrstiti med nepopravljeno (najmanj berljivo) besedilo in popolnoma popravljeno (najbolj berljivo) besedilo.

5. Metoda

5.1. Udeleženci

V preizkušnji je sodelovalo 33 udeležencev, za končno analizo pa smo uporabili rezultate 31 udeležencev (26 žensk). Povprečna starost je bila 20,2 leti (*SD* = 2,4), najnižja 18 let, najvišja pa 27 let. Udeleženci so bili dodiplomski oziroma podiplomski študenti na Univerzi v Ljubljani in so bili materni govorci slovenščine. Vsi so imeli normalen vid ali pa so pri branju uporabljali korekcijo vida (leče ali očala).

5.2. Oprema

Celotna preizkušnja je bila vodena računalniško. Za snemanje smo uporabljali sistem EyeLink 1000 (frekvenca zbiranja podatkov: 1000 Hz) v psihološkem laboratoriju Filozofske fakultete Univerze v Ljubljani. Da bi dosegli čim boljšo kvaliteto podatkov, so imeli med branjem udeleženci glavo naslonjeno na posebnem stojalu, ki je bilo od zaslona oddaljeno približno 60 cm. V vseh primerih smo merili gibanje desnega očesa. Uporabljali smo 9-točkovni kalibracijski model.

5.3. Besedila

Povprečna dolžina vseh besedil, uporabljenih v preizkušnji je znašala 223 besed (*SD* = 7,2) oziroma 1200 znakov brez presledkov (*SD* = 11,1). Vsa, razen enega besedila v vaji, so bila preoblikovana v dva odstavka (pisava Calibri, velikost 12, 1,5 medvrstični razmik), naslov je bil vedno v odebeljenem tisku.

Vsa besedila, razen enega v vaji, smo dobili na spletni strani Evropske komisije. Glavni razlog za takšno odločitev ni vsebinske, pač pa v prvi vrsti predvsem pragmatične narave; vse angleške novice na spletni strani so namreč prevedene tudi v slovenščino. V primeru zagat pri pripravljanju oziroma popravljanju besedilnih različic za preizkušnjo smo imeli tako na voljo ustrezno referenco. Poleg tega se vse novice na strani Evropske komisije dotikajo Evropske unije, s čimer smo zagotovili vsebinsko primerljivost vsaj na najvišji besedilni ravni.

5.4. Preizkusne različice

Za vsako od treh besedil, uporabljenih v preizkušnji, smo pripravili tri različice, ki so v našem primeru ustrezale trem eksperimentalnim pogojem. V izhodišču smo angleška besedila samo prevedli s pomočjo Googlovega prevajalnika. To različico smo označili z NP (nepopravljena različica). V naslednjem koraku smo v vseh strojnih prevodih popravili zgolj ujemalne napake, s čimer smo ustvarili delno popravljene različice, ki smo jih označili s P1 (popravljena različica 1). Nazadnje smo v vseh besedilih popravili še preostale napake in tako dobili popolnoma popravljena besedila. Te smo označili s P2 (popravljena različica 2). Na koncu smo dobili skupno 9 možnih kombinacij besedila (3) in tipa prevoda (3). Vsi predstavljeni rezultati za posamezen tip prevoda so nato združeni preko različnih besedil.

5.5. Potek preizkušnje

Vsem udeležencem je bilo podano enako navodilo, da naj berejo za razumevanje in na način, ki jim najbolj ustreza. Posebej je bilo poudarjeno, da si podatkov v besedilih ni potrebno dodatno zapomniti ter da se v nadaljevanju preizkušnje uporabljeni vprašalniki ne nanašajo na poznavanje vsebine prebranega. Ko so prebrali navodila, so udeleženci s pritiskom na odgovorno tablico sprožili prikaz besedila in začetek snemanja očesnih gibov. Konec branja so prav tako označili s pritiskom na gumb, s čimer se je zaključilo tudi snemanje. Pri branju ni bilo časovne omejitve. Posamezna preizkušnja, ki je zajemala branje petih besedil in odgovarjanje na vprašanja, je praviloma trajala med 20 in 30 minut.

5.6. Subjektivne ocene in verbalna poročila

Poleg mer sledilca očesnih gibov smo želeli med preizkušnjo pridobiti tudi subjektivno oceno branja vsakega posameznika. Po vsakem prebranem besedilu so udeleženci z uporabo kontinuirane lestvice (z miško so morali klikniti na ustrezni del prikazane lestvice oziroma skale) odgovorili na tri vprašanja o poteku branja, zahtevnosti razumevanja in vračanju na že prebrane dele besedila oziroma t. i. »regresijah«.

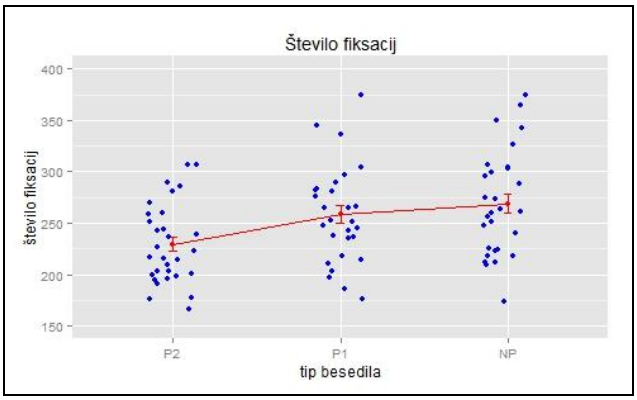
Po končanem branju smo z vsakim udeležencem opravili še krajši pogovor, da bi pridobili čim več dodatnih informacij, ki bi lahko koristile pri interpretaciji rezultatov sledilca očesnih gibov in subjektivnih ocen. Zastavili smo jim naslednja vprašanja: a) katero besedilo je bilo za branje najtežje in katero najlažje, b) ali bi bili sposobni povzeti vsebino prebranega, c) ali so med branjem spremenili bralno strategijo in d) ali obstaja kakšen tip napake, ki so si ga posebej zapomnili.

6. Rezultati

6.1. Sledilec očesnih gibov

6.1.1. Število fiksacij

Število fiksacij predstavlja število vseh fiksacij, narejenih med branjem posameznega besedila. Fiksacija je bila po vzoru predhodnih študij definirana kot vsak postanek, daljši od 100 ms.

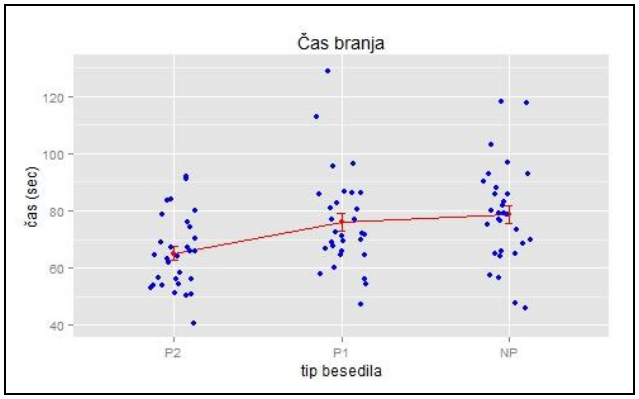


Slika 1: Število fiksacij

Enosmerna analiza variance (ANOVA) pokaže, da so razlike statistično pomembne glede na tip prevoda, $F(2, 60) = 16{,}9$, $p < 0{,}05$, in v manjši meri tudi glede na posamezno besedilo, $F(2, 60) = 3{,}4$, $p < 0{,}05$. Ločeni t-test pokaže statistično pomembne razlike med popolnoma popravljeno in nepopravljeno vrsto prevoda, $t(30) = 5{,}3$, $p < 0{,}05$. Med delno popravljeno in nepopravljeno različico ni bilo statistično pomembnih razlik, $t(30) = 1{,}2$, $p > 0{,}05$.

6.1.2. Čas branja

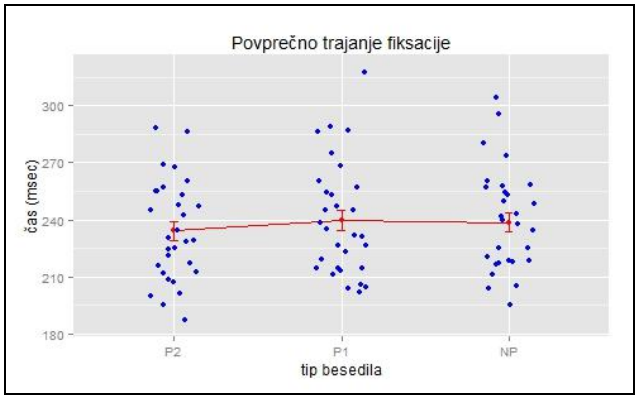
Čas branja je seštevek trajanja vseh fiksacij, narejenih med branjem besedil.



Slika 2: Čas branja

Enosmerna analiza variance (ANOVA) pokaže, da so razlike statistično pomembne glede na tip prevoda, $F(2, 60) = 18{,}4$, $p < 0{,}05$, in v manjši meri tudi glede na posamezno besedilo, $F(2, 60) = 3{,}1$, $p < 0{,}05$. Ločeni t-test pokaže statistično pomembne razlike med popolnoma popravljeno in nepopravljeno vrsto prevoda, $t(30) = 6$, $p < 0{,}05$. Med delno popravljeno in nepopravljeno različico ni bilo statistično pomembnih razlik, $t(30) = 1{,}2$, $p > 0{,}05$.

6.1.3. Povprečno trajanje fiksacije

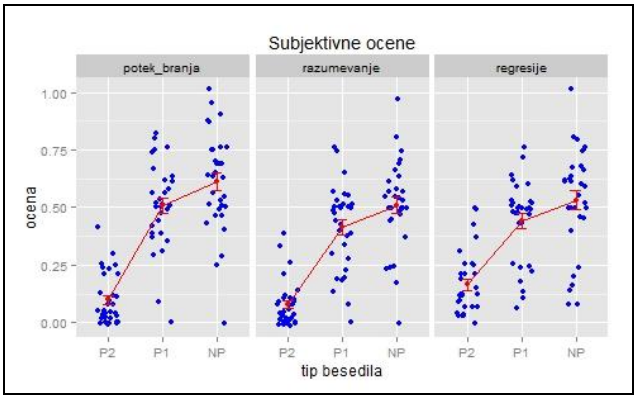


Slika 3: Povprečno trajanje fiksacije

Enosmerna analiza variance (ANOVA) pokaže, da so razlike statistično pomembne glede na tip prevoda, $F(2, 60) = 4{,}7$, $p < 0{,}05$. Glede na posamezno besedilo razlike niso bile statistično pomembne, $F(2, 60) = 0{,}6$, $p > 0{,}05$. Ločeni t-test pokaže statistično pomembne razlike med popolnoma popravljeno in nepopravljeno vrsto prevoda, $t(30) = 2{,}6$, $p < 0{,}05$. Med delno popravljeno in nepopravljeno različico ni bilo statistično pomembnih razlik, $t(30) = 0{,}7$, $p > 0{,}05$.

6.2. Subjektivne ocene

Odgovori, zbrani s kontinuiranimi lestvicami, so predstavljeni kot proporci lestvice, pri čemer nižje vrednosti (bližje 0) predstavljajo pozitivno vrednotenje (na primer: »branje je potekalo povsem tekoče«), medtem ko višje vrednosti proti 1 predstavljajo negativno vrednotenje (na primer »razumevanje je bilo zelo oteženo«).



Slika 4: Subjektivne ocene

Enosmerna analiza variance (ANOVA) za kriterij berljivosti pokaže, da so razlike statistično pomembne glede na tip prevoda, $F(2, 60) = 112$, $p < 0{,}05$. Ločeni Friedmanov test pokaže statistično pomembne razlike med popolnoma popravljeno in nepopravljeno vrsto prevoda, $\chi^2 = 30$, $p < 0{,}05$. Razlike v ocenah so bile statistično pomembne tudi med delno popravljenimi in nepopravljenimi prevodi , $\chi^2 = 5{,}1$, $p < 0{,}05$.

Ne glede na kriterij ocenjevanja, po katerem se je spraševalo (potek branja, razumevanje ali regresije), opažamo razlike v povprečni oceni za posamezen tip besedila in enak trend naraščanja povprečnih vrednosti. Povsem popravljena besedila (P2) so bila po vseh treh kriterijih ocenjena z nižjimi vrednostmi kot delno popravljena besedila (P1) in nepopravljena besedila (NP).

6.3. Korelacije

Zanimalo nas je tudi, ali morda obstaja povezanost med subjektivnimi ocenami udeležencev in merami sledilca očesnih gibov. V tabeli 2 so predstavljeni koeficienti korelacije med subjektivnimi ocenami berljivosti in tremi merami sledilca. Vrednosti v odebeljenem tisku so statistično pomembne pri kritični meji $p = 0{,}05$.

	ocena ~ št. fiksacij	ocena ~ čas branja	ocena ~ pov. čas fiksacije
besedilo 1	0,26	**0,40**	**0,46**
besedilo 2	**0,43**	**0,41**	0,15
besedilo 3	**0,51**	**0,52**	0,16

Tabela 2: Koeficienti korelacije (*r*) pri posameznih besedilih

6.4. Verbalna poročila

Tabela 3 povzema število odgovorov za posamezen tip besedila. Udeleženci so morali povedati, katero od treh prebranih besedil je bilo najtežje in katero najlažje, pri čemer je bilo možno ostati neodločen.

V tabeli 3 so zajeti le odgovori 22 udeležencev, ker preostalim nismo zastavili povsem identičnih vprašanj.

Nekateri so morali namreč povedati le, katero besedilo je bilo najtežje. Med temi so trije kot najtežje izbrali besedilo iz pogoja NP, eden P2, eden pa je ostal neodločen. Druge odgovore smo izključili, ker smo po naknadnem preverjanju opazili, da so udeleženci izbirali besedila iz vaje in je šlo torej za napačno razumevanje vprašanja, ki se je nanašalo le na zadnja 3 besedila.

	NP	P1	P2	nobeno
najtežje branje	17	3	0	2
najlažje branje	0	1	19	2

Tabela 3: Verbalno poročanje o berljivosti (N = 22)

Na vprašanje, ali bi bili sposobni po branju povzeti najosnovnejše informacije o prebranih besedilih, je 6 udeležencev odgovorilo z »ne«, 25 pa z »da«. Na vprašanje, ali so v primeru, ko so v besedilih naleteli na napake, na kakršenkoli način spremenili način branja oziroma bralno strategijo, je 22 udeležencev odgovorilo pritrdilno, 9 pa je bilo mnenja, da načina branja niso spreminjali. Pri tem so udeleženci omenjali obe možni spremembi branja – nekateri so trdili, da so začeli brati počasneje z vračanjem na že prebrane dele besedila, medtem ko je kar nekaj udeležencev potrdilo, da se niso posebej posvečali napakam v besedilu in so posledično brali celo malce hitreje kot običajno. Na vprašanje, ali so si kakšen tip napake posebej zapomnili, eden od udeležencev ni izpostavil ničesar, medtem ko so vsi ostali lahko našteli vsaj en tip napake. Najpogosteje so navajali napačen besedni red, neujemanje (»napačen sklon«, »napačen spol«, »napačne oblike besed«, »nedoločnik, kjer ni potrebno«), napačno postavljene vejice ter vrinjene besede (»ponavljanje besed«).

7. Diskusija

7.1. Primerjava rezultatov

V povprečju so bila nepopravljena besedila fiksirana bolj pogosto, čas branja pa je bil daljši kot pri delno popravljenih in povsem popravljenih različicah, kar je v skladu z začetno hipotezo. Do statistično pomembnih razlik je kljub temu prišlo le med obema skrajnima pogojema, ne pa tudi med delno popravljenimi in nepopravljenimi prevodi.

Takšne rezultate lahko razumemo kot delno skladne z ugotovitvami Dohertyja in soavtorjev (2010), kjer so udeleženci brali zelo dobro prevedene in zelo slabo prevedene stavke, učinek pa je bil razviden v številu fiksacij in času branja. Po drugi strani pa so rezultati za povprečno trajanje fiksacije nepričakovani in jih z našo hipotezo težko učinkovito pojasnimo. Na prvi pogled so bile razlike med tremi pogoji majhne, čeprav analiza variance razkrije sistematično razliko med posameznimi tipi prevodov. Glede na to, da se ti rezultati ne skladajo s predhodno študijo, iščemo morebitno razlago v dejstvu, da smo sami uporabljali cela besedila, medtem ko so v študiji Dohertyja in soavtorjev (2010) uporabljali le posamezne stavke.

Naslonili bi se lahko namreč na fenomen »sinteze« (ang. *wrap up effects*). V več študijah se je pokazalo, da je trajanje fiksacije na zadnjih besedah oziroma ob koncu stavka daljše kot v predhodnih delih. Raziskovalci so si ta pojav razlagali s hipotetičnim procesom sinteze prebrane vsebine v smiselno celoto, ki poteče na koncu prebrane enote (Staub et al., 2007). Možna razlaga naših rezultatov bi tako bila, da pride med branjem besedil večkrat do integracije prebranega dela v celoto. Ker je v slabih, manj berljivih besedilih takšna sinteza bolj potrebna, lahko do razlike v povprečnem trajanju fiksacij pride le na ravni besedil, ne pa tudi na ravni posameznih stavkov.

Bolj kot rezultati specifičnih mer sledilca pa so za naš namen zanimive primerjave s človeškimi ocenami in verbalnimi poročili. Ne glede na kriterij, po katerem smo spraševali, je bil trend subjektivnih ocen enak: popravljene različice so v povprečju dobivale višje ocene kot delno popravljene, te pa so bile ocenjene boljše kot nepopravljeni prevodi. Ocene med drugim jasno kažejo, da so udeleženci vse tri tipe prevodov, ne le najskrajnejše, dojemali kot kvalitativno različne. Podobno je moč sklepati tudi iz rezultatov verbalnih poročil, ki kažejo na to, da udeleženci niso (kvalitativno) enačili delno popravljenih in nepopravljenih prevodov. Na splošno so bile ocene bolj enotne za povsem popravljene različice kot za delno in nepopravljene različice, kjer so ocene bolj razpršene. Podobno kot pri Stymne in soavtorjih (2012) tudi naši rezultati sicer pokažejo zmerno korelacijo z ocenami udeležencev, pri čemer je bila korelacija najbolj konsistentna za čas branja, najmanj pa za povprečno trajanje fiksacije.

7.2. Sledilec in evalvacija strojnih prevodov

Neskladje med različnimi viri podatkov za vrednotenje berljivosti je zanimivo. Glede na splošno sprejeto prakso pri evalvaciji strojnih prevodov vzamemo rezultate človeških ocen kot standard za ovrednotenje rezultatov sledilca, te pa pokažejo, da so udeleženci razlikovali med vsemi tipi strojnih prevodov, uporabljenih v preizkušnji.

Menimo, da je inherentna omejitev uporabe sledilca za predlagani namen ta, da rezultati (kjer je variabilnost med posamezniki precejšnja) ne odsevajo zgolj kvalitete besedila, temveč nujno zajemajo tudi druge vplive na kognitivno procesiranje: posameznikovo bralno kompetenco, slog branja, odziv na navodila v nalogi, tip besedila ipd. Posameznik tako lahko prepozna določeno besedilo kot manj kvalitetno, vendar to zaradi množice potencialnih dejavnikov na branje morda ne bo nujno razvidno iz opazovanega vedenja, tj. gibanja oči med branjem.

Če na podlagi mer sledilca res ni možno razlikovati med takšnimi besedili, ki ne vsebujejo precejšnjega dela napak, in tistimi besedili, ki vsebujejo prav vse napake, potem smo lahko skeptični do smotrnosti uporabe sledilca v evalvacijske namene. Predvidevamo namreč, da so v realnih okoliščinah, tj. pri razvoju novih prevajalnikov, razlike v prevodih nekega sistema in v prevodih nekoliko izboljšane različice tega istega sistema lahko precej manj očitne, kot so bile razlike v besedilih iz naše preizkušnje. Vprašanje, ali je besedilo brez napak bolj berljivo od takega z veliko napakami pa smatramo za trivialno in kot tako ne zahteva preverjanja s tako zahtevno metodologijo. Ključno vprašanje je namreč, ali lahko na podlagi

rezultatov sledilca dobimo informacijo, ki nam je druge metode ne omogočajo.

Glede na to, da, kolikor nam je znano, sistematično zbrani podatki sledilca očesnih gibov med branjem za slovenščino še niso na voljo (prim. Ferbežar, 2012), bi se vzdržali prestrogega vrednotenja metodologije. Vsakršno nadaljnje delo na tem področju bi pripomoglo tudi k boljši kvaliteti podatkov, ki v našem primeru ni bila vedno zadovoljiva.

V morebitnih nadaljnjih študijah bi bilo zanimivo opraviti analizo podatkov na ravni posameznih besed, kot so to storili Stymne in soavtorji (2012), kjer se je prav ta pristop izkazal za bolj informativnega od uveljavljenih tehnik. V naši študiji zaradi tehničnih težav na koncu ta možnost ni bila izvedljiva. Poleg analize na ravni besed, bi lahko v morebitnih prihodnjih študijah tako testirali še druge, bolj specifične mere sledilca, kot je npr. število regresij (sakade na že prebrane dele besedila), ki se pogosto povezujejo s težavami razumevanja med branjem.

Nadaljnja raziskovalna vprašanja se odpirajo tudi pri uporabi kontinuiranih lestvic za ocenjevanje, namesto uveljavljenih 5- ali 7-stopenjskih (prim. Graham et al., 2013a).

8. Zaključek

Z rezultati sledilca očesnih gibov smo lahko le delno potrdili uvodno hipotezo. Mere sicer pokažejo pričakovan trend naraščanja, a se ne izkažejo kot zelo učinkovit pokazatelj razlik med posameznimi tipi prevodov. Na podlagi rezultatov te študije tako menimo, da sledilec očesnih gibov našega testa ni prestal: da bi ga lahko označili kot uporabno orodje za evalvacijo, bi morali rezultati pokazati sliko, skladno s predvidevanjem na podlagi kategorizacije napak in ocenami udeležencev.

Kljub sicer nekoliko bolj skeptičnemu zaključku smo mnenja, da gre za obetaven pristop, ki sicer zahteva kar nekaj metodološke pozornosti in natančnosti, a hkrati ponuja drugačen vpogled v obravnavano tematiko. V prihodnje bi sledilec očesnih gibov ob ustrezno izboljšani kvaliteti podatkov veljalo izkoristiti za bolj podrobno analizo branja ter tudi za bazične raziskave branja.

9. Literatura

Callison-Burch, C., F. Cameron, P. Koehn, C. Monz in J. Schroeder, 2007. (Meta-) Evaluation of Machine Translation. *Proceedings of ACL-2007 Workshop on Statistical Machine Translation*.

Doherty, S., S. O'Brien in M. Carl, 2010. Eye-tracking as MT evaluation technique. *Machine Translation*, 1: 1–13.

Ferbežar, I., 2012. *Razumevanje in razumljivost besedil*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Graham, Y., T. Baldwin, A. Moffat in J. Zobel, 2013a. Continuous Measurment Scales in Human Evaluation of Machine Translation. In *Linguistic Annotation Workshop and Interoperability with Discourse*, 33–41. 8. in 9. avgust, Sofija, Bolgarija.

Graham, Y., T. Baldwin, A. Moffat in J. Zobel, 2013b: Crowd-Sourcing of Human Judgments of Machine Translation Fluency. *Proceedings of the Autralasian Language Technology Workshop*, 16–24. 4.–6. december, Dunedin, Nova Zelandija.

Just, M.A. in P.A. Carpenter, 1980. A Theory of Reading: From Eye Fixations to Comprehension. *Psychological Review*. 4:329-354.

Koehn, P., 2010. *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Papineni, K., S. Roukos, T. Ward in W.-J. Zhu, 2002: BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Julij 2002, Philadelphia, ZDA.

Rayner, K., 1998. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, 3: 372–422.

Staub, A. in K. Rayner, 2007. Eye Movements and on-line comprehension processes. Gaskell, G. (ur.), *The Oxford Handbook of Psycholinguistics*, Oxford: Oxford University Press.

Stymne, S., H. Danielsson, S. Bremin, H. Hu, J. Karlsson, A.P. Lillkull in M. Wester, 2012. Eye Tracking as a Tool for Machine Translation Error Analysis. *Zbornik konference LREC 2012*, 1121–1126.

Verdonik, D. in M. Sepesy Maučec, 2013. O avtomatski evalvaciji strojnega prevajanja. *Slovenščina 2.0.* 1/1. 111–113.

Evalvacija slovensko-srbskih strojnih prevodov v projektu SUMAT

Mirjam Sepesy Maučec

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko
Smetanova 17, SI-2000 Maribor
mirjam.sepesy@um.si

Povzetek

V prispevku predstavljamo postopek in rezultate evalvacije prevodov statističnega strojnega prevajalnika podnapisov, ki smo ga razvili v okviru evropskega projekta SUMAT. Cilj projekta je bil razviti orodje za strojno prevajanje podnapisov, ki bi olajšalo delo profesionalnih prevajalcev. V projektu smo predlagali uporabo strojnih prevodov kot osnovo za kvalitetno prevajanje, ki vključuje tudi popravljanje strojnih prevodov, ki ga opravi profesionalni prevajalec. Uporaba strojnih prevodov je smiselna, če prevajalcu olajša delo oz. mu prihrani čas, zato je bila ključna naloga v projektu evalvacija prevodov, tako v smislu njihove kvalitete kot prihranka časa. V prispevku prikažemo rezultate evalvacije za prevajanje iz slovenščine v srbščino. Ta je pokazala, da je lahko sistem SUMAT učinkovito orodje za prevajalce podnapisov. V projektu je bila opravljena tudi evalvacija za prevajanje v obratni smeri, ki je dala podobne rezultate, saj gre za prevajanje med dvema sorodnima jezikoma.

Evaluation of machine translation for Slovenian – Serbian in the project SUMAT

This article describes the evaluation of statistical machine translation as carried out during the SUMAT project. The goal of this project was to build a tool for the automatic translation of subtitles that would help professional translators. Machine translation is useful if it makes subtitler's job easier and saves him/her time. The idea of the project was to use the translations obtained by the tool as the basis for post-editing, which should be done by professional translators in order to obtain translations of required quality. The crucial part in the project was the evaluation of machine translations quality, and the measurement of productivity gain/loss. The results for Slovenian – Serbian translation show that SUMAT system could be a useful tool for professional translators. This article only presents those results for translation from Slovenian to Serbian are presented. Similar results were however obtained for translation in opposite direction.

1. Uvod

Danes je večina besedilnega gradiva na voljo v elektronski obliki. To daje korpusnim in statističnim pristopom v jezikovnih tehnologijah veliko prednost, tudi na področju strojnega prevajanja. Statistično strojno prevajanje se je skozi številne raziskave pokazalo kot najučinkovitejši pristop k avtomatskemu prevajanju. Dodaten razlog za njegov uspeh je tudi ta, da za razvoj prevajalnika ni potrebno poglobljeno znanje o jezikih, med katerimi prevajamo.

Zahtevnost strojnega prevajanja je odvisna od žanra in domene besedil, ki jih prevajamo. Sprva je kazalo, da je prevajanje podnapisov, s katerim smo se ukvarjali v projektu SUMAT, za statistično strojno prevajanje zelo hvaležno področje, saj so povedi praviloma kratke. Toda podnapisi prinašajo tudi številne probleme. Ker gre za prevajanje podnapisov video vsebin, so nekateri problemi blizu problemom govorjenega jezika. Še večji problem pa je, da so se mora dolžina besedila podrejati dolžinam podnapisa, kar privede do številnih postopkov krajšanja izvornega besedila.

V projektu SUMAT smo razvili statistični strojni prevajalnik za prevajanje podnapisov med 14 jezikovnimi pari, med temi je tudi za slovenščino-srbščino, ki ga obravnavamo v tem članku. Ideja projekta je bila uporabiti že razvite metode statističnega strojnega prevajanja in zgraditi prevajalsko orodje, ki bo v pomoč profesionalnim prevajalcem pri generiranju kvalitetnih prevodov. Cilj ni bil, da bi prevajalnik tvoril brezhibne prevode, ampak da bi prevajalcu ponudil prevod, ki ga bo le-ta s čim manj dela preoblikoval v prevod željene oz. zahtevane kakovosti.

V projektu so bili uporabljeni prevodi profesionalnih prevajalcev, ki so v lasti prevajalskih podjetij, partnerjev v projektu in izven projekta niso dostopni.

2. Korpus SUMAT

Osnovno gradivo statističnega prevajalnika predstavlja vzporedni korpus. Od kvalitete vzporednega korpusa je neposredno odvisna uspešnost prevajanja, saj je uporabljen kot učni korpus prevajalnika. V projektu SUMAT so gradivo za vzporedni korpus iz svojih arhivov prispevala mednarodna podjetja, ki se profesionalno ukvarjajo s prevajanjem podnapisov. Izvorno gradivo ni neposredno uporabno, ampak ga je treba obdelati. Koraki procesiranja so: pretvorba v enoten format; poravnavanje dokumentov; tokenizacija; poravnavanje podnapisov in normalizacija (zapis z malimi črkami). Procesiranje korpusa SUMAT za slovensko-srbski jezikovni par je podrobneje predstavljeno v Maučec et al. (2012). Vzporedni korpus obsega 167.700 poravnanih podnapisov, kar je okrog 1,7 mio. besed v slovenskem jeziku in 2 mio. besed v srbskem.

Korpus takega obsega je za gradnjo »uporabnega« prevajalnika premajhen, zlasti pri visoko pregibnih jezikih, kot sta tako slovenščina kot srbščina; za tovrstne jezikovne pare potrebujemo za doseganje zadovoljive pokritosti besedišča čim večje korpuse, vsaj 10 mio. besed ali več. V projektu smo zato učni korpus dopolnili še z neprečiščenim gradivom iz prosto dostopnega korpusa OpenSubtitles (Tiedemann, 2009) (1,9 mio poravnanih podnapisov) in z interno zbirko prevodov popularnih filmov (44.500 podnapisov). Celoten korpus, ki smo ga uporabili za učenje prevajalnika, je vseboval 2,1 mio. podnapisov (16,8 mio besed za slovenski in 17,6 mio besed za srbski jezik).

Nepogrešljiva komponenta prevajalnika je tudi jezikovni model. Za njegovo učenje uporabimo enojezično gradivo. V projektu smo uporabili vse razpoložljivo gradivo iz prej omenjenih virov, tj. tudi tiste segmente, ki jih v pripravi vzporednega korpusa nismo uspeli poravnati. Slovenski korpus je obsegal 4,35 mio. podnapisov oz. 36 mio. besed, srbski korpus pa 4,56 mio. podnapisov oz. 42 mio. besed.

Slovar prevajalnika (tj. besede, ki jih prevajalnik prevaja) smo izluščili iz vzporednega korpusa. Tako je slovenski slovar vseboval 394.000 besed, srbski pa 570.000 besed.

3. Gradnja prevajalnika SUMAT

Prevajalnik SUMAT ima klasično strukturo. Kot osnovna enota prevajanja se običajno uporablja poved, v projektu pa je bilo opravljenih nekaj preliminarnih testov, ki so vodili v odločitev, da kot osnovno enoto uporabimo podnapis.

Prevajalnik sestavljajo 3 osnovne komponente: model prevajanja, model preurejanja in jezikovni model. Prvi dve komponenti smo zgradili s pomočjo orodja Moses (Koehn et al., 2007), jezikovni model pa z orodjem SRI LM (Stolcke, 2002).

Učni vzporedni korpus izhaja iz treh različnih virov, zato smo modela prevajanja in preurejanja gradili za vsak vir posebej in jih potem sestavili po principu adaptacije na domeno (Sennrich, 2012). Kot vzorec ciljne domene smo uporabili razvojno množico, ki je obsegala 2000 podnapisov.

Za izgradnjo jezikovnih modelov smo uporabili celotni korpus podnapisov. Uporabili smo 3-gramski jezikovni model z Good-Turingovim odštevanjem in sestopanjem po Katz. Perpleksnost slovenskega jezikovnega modela na testni množici je znašala 206, na srbski pa 230. Preizkusili smo tudi dodajanje gradiva iz enojezičnega korpusa pisanega jezika, ki rezultata ni izboljšalo.

Uteži komponent prevajalnika smo optimirali po MERT (Och, 2003) na razvojni množici 2000 podnapisov.

4. Evalvacija

V prvem delu smo izvedli avtomatsko evalvacijo s 4000 naključno izločenimi podnapisi, ki niso bili ročno pregledani. Uporabili smo metrike avtomatske evalvacije BLEU in TER (Papineni et al., 2002; Snover et al, 2006). Zanimal nas je tudi delež podnapisov, ki se 100% ujemajo z referenco (Equal), in delež podnapisov, pri katerih je, da dosežemo ujemanje, potrebnih največ 5 korakov preurejanja (Lev5). Rezultati evalvacije so v prvi vrstici v tabeli 1. Rezultati so slabi. Vzrok za to je v nastanku slovensko-srbskih SUMAT poravnanih podnapisov. Le-ti niso bili generirani kot neposredni slovensko – srbski prevodi, ampak oboji neposredno iz video signala v angleščini. To je prevajalcem iz angleščine v srbščino in slovenščino ponujalo veliko mero svobodne pri izbiri besed. Testno množico SUMAT smo podrobneje analizirali v (Verdonik & Maučec, 2013).

	BLEU	TER	Equal	Lev 5
Testna	17,80	66,10	4,00	11,60
Faza 1	46,30	36,20	13,80	38,90
Faza 2	57,40	26,30	22,10	47,90
Faza 3	69,20	17,30	37,40	69,10
SUMAT povprečje	39,69	44,88	20,1	35,69

Tabela 1: Rezultati evalvacije v različnih fazah projekta

V drugem delu evalvacije so bili v ocenjevanje kvalitete prevodov vključeni profesionalni prevajalci. Evalvacija s prevajalci je potekala v dveh sklopih. V prvem sklopu smo z njihovo pomočjo izboljševali sistem, v drugem delu pa merili, ali se učinkovitost prevajanja ob uporabi sistema izboljša, če se torej čas prevajanja z uporabo strojnih prevodov kaj skrajša.

4.1. Ocenjevanje s pomočjo prevajalcev

Prvi sklop evalvacije je potekal v treh fazah. Za vsako fazo je bila izbrana datoteka, ki je predstavljala zaključeno celoto, npr. podnapise celotnega filma, dokumentarca, pogovorne oddaje ipd. Datoteka je bila prevedena s pomočjo sistema in dana prevajalcu v pregled. Pregled je vključeval: rangiranje prevoda glede na kvaliteto, klasifikacijo napak in popravljanje prevoda v pravilnega. Prevajalci so v vsaki fazi v posebni datoteki podali tudi predloge popravkov.

4.1.1. Avtomatska evalvacija z ročno popravljenimi strojnimi prevodi za referenco

Strojne prevode dokumentov vseh treh faz smo avtomatsko evalvirali, tako da smo kot referenco uporabili popravljene prevode, ki so jih zapisali prevajalci. Rezultati so zbrani v vrsticah od 2 do 4 v tabeli 1. Vidimo, da so rezultati neprimerno boljši kot v primeru testne množice SUMAT. Razvidno je tudi, da so se rezultati iz faze v fazo izboljševali. V zadnji vrstici v tabeli 1 so povprečni rezultati avtomatskih metrik po vseh jezikovnih parih projekta. Tudi iz te primerjave lahko razberemo, da so bili v primerjavi z drugimi jezikovnimi pari za prevajanje slovenščina-srbščina doseženi zelo dobri rezultati. To je pričakovano, saj gre za sorodna jezika.

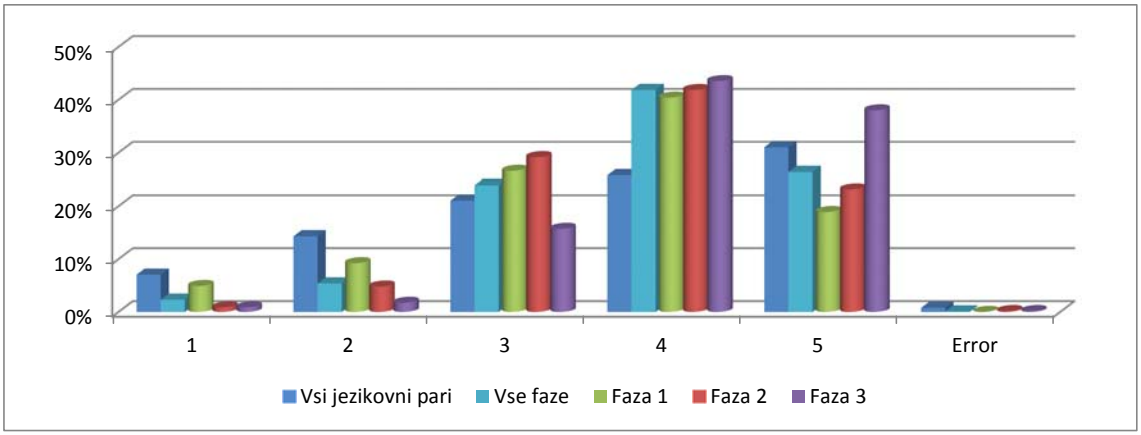
4.1.2. Rangiranje prevodov

Prevajalci so vsak podnapis v strojnem prevodu rangirali glede na kvaliteto oz. zahtevnost popravljanja. Pri tem smo uporabili skalo, definirano v "WMT 2012 Shared Task on MT quality estimation", po kateri je vsak podnapis rangiran z vrednostjo od 1 do 5. Ocena 1 pomeni neuporaben in nerazumljiv prevod, ocena 5 pa brezhiben prevod, ki ne potrebuje nobenega popravka. Rezultati za prevode slovenščina-srbščina so zbrani v tabeli 2. Vidimo, da je največ prevodov dobilo oceno 4. Več kot 20 % prevodov ima oceno 5, kar pomeni, da je petina prevodov neposredno uporabnih. Tudi iz te tabele je razvidno, da so se rezultati iz faze v fazo izboljševali.

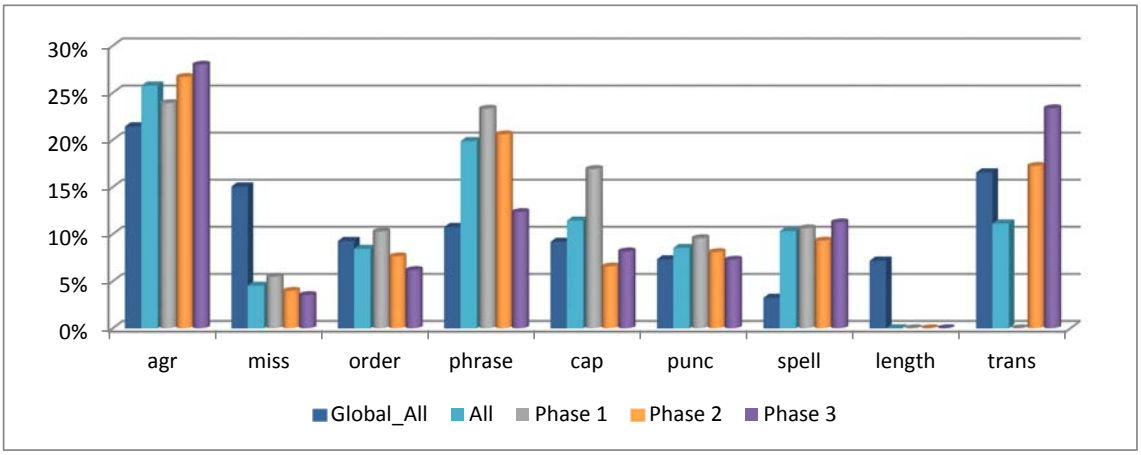
4.1.3. Klasifikacija napak

Prevajalci so napake v prevodih klasificirali v razrede:

- *agr*: slovnično neujemanje,
- *miss*: manjka polnopomenska beseda ali odsek,
- *order*: napačni vrstni red besed,
- *phrase*: večbesedna zveza napačno prevedena kot ločene, nepovezane besede,
- *cap*: napačen zapis velike/male črke,
- *punc*: napačno ločilo,
- *spell*: napačno črkovanje,



Slika 1: Rangiranje prevodov glede na oceno kvalitete



Slika 2: Deleži napak v različnih fazah projekta

- *length*: predolg prevod glede na omejeno dolžino podnapisa,
- *trans*: napačen prevod.

Slika 1 prikazuje deleže napak po fazah. Vidimo lahko, da se je delež nekaterih napak skozi faze manjšal (npr. *cap*, *punct*, *phrase*, *order*), nekaterih pa celo povečal (npr. *trans*).

4.1.4. Popravki v sistemu

Na osnovi klasifikacije napak in predlogov prevajalcev smo sistemu dodali nekaj korakov naknadne obdelave strojnih prevodov. Glede na končna ločila smo popravili velike začetnice besed. Dodali smo nekaj 100 pravil za popravljanje slovničnega neujemanja. Definirali smo tudi nekaj pravil za zapis števil. Brisali smo presledke pred ločili.

Napačnih prevodov nismo uspeli popravljati. Vzrok za nekatere napačne prevode je neupoštevanje konteksta. Prevajalnik obravnava podnapis kot zaključeno celoto, ne glede na vsebino predhodnih podnapisov.

Reševanje določenih napak je pogojeno z uporabo dodatnih jezikovnih virov, ki jih zaradi komercialne naravnanosti projekta nismo dodajali, saj je za vsak uporabljen vir potrebno dovoljenje za komercialno rabo.

Nekatere napake smo odpravili tudi s tem, da smo dodali še en korak optimizacije uteži z MERT, tako da smo razvojno množico nadomestili z datotekami iz drugega dela evalvacije.

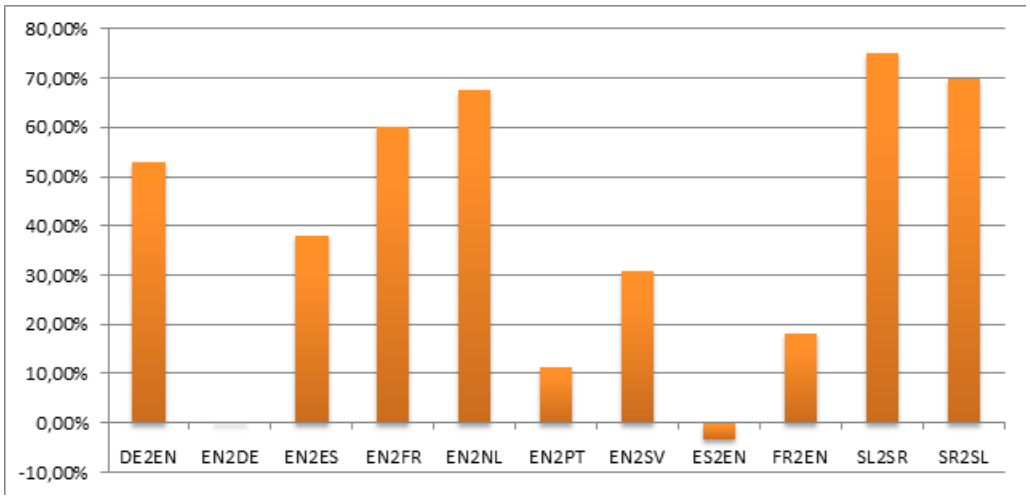
V fazi izboljševanja sistema učnega korpusa nismo spreminjali, čeprav smo ugotovili, da OpenSubtitles korpus vsebuje veliko šuma. V okviru projekta bi bilo čiščenje korpusa časovno preveč zahtevno.

4.2. Merjenje produktivnosti

V drugem delu evalvacije s prevajalci smo merili učinkovitost uporabe strojnih prevodov, ki jih je generiral izboljšan sistem. Primerjali smo čas, ki ga potrebuje prevajalec, če neposredno prevaja dokument iz izvornega v ciljni jezik, s časom, ki ga potrebuje za naknadno obdelavo strojnih prevodov. Menimo, da je tovrstna primerjava zelo jasen in neposreden pokazatelj uporabnosti sistemov strojnega prevajanja.

Pred izvedbo drugega dela evalvacije smo v sistem prevajanja vpeljali še dodaten postopek filtriranja strojnih prevodov. V razdelku 5.1 smo opisali rangiranje prevodov glede na kvaliteto. Na osnovi teh ocen smo učili binarni klasifikator, ki prevode klasificira v dva razreda, v razred dobrih in razred slabih prevodov. Za učenje klasifikatorja in klasifikacijo smo uporabili orodje QuEst, ki je podrobneje opisano v (Specia et al., 2013). Strojne prevode, ki jih je klasifikator označil kot slabe, smo odstranili, kar je pomenilo, da jih mora prevajalec tvoriti iz podnapisa v izvornem jeziku.

Za vsak jezikovni par oz. za vsako smer prevajanja sta sodelovala dva profesionalna prevajalca. Izjema je jezikovni par slovenščina-srbščina, kjer je za vsako smer prevajanja sodeloval le en prevajalec. Vsak prevajalec je tvoril tri datoteke. V prvi je prevajal iz izvornega jezika, v drugi je popravljal strojne prevode in v tretji je popravljal le filtrirane strojne prevode. Pri tem je vsak prevajalec



Slika 3: Rast produktivnosti pri uporabi strojnih prevodov v prevajalskem procesu

uporabil programsko okolje, ki ga tudi sicer uporablja pri svojem delu. Razlika je bila le v tem, da se je v ozadju meril čas efektivnega dela.

Iz primerjave časov, potrebnih za generiranje prevodov v prvi in drugi datoteki oz. prvi in tretji datoteki, smo izračunali rast/padec produktivnosti (ang. productivity gain/loss) pri prevajalskem procesu. Rezultati učinkovitosti uporabe strojnih prevodov za vse jezike v projektu SUMAT so prikazani na sliki 2. Vidimo, da je pri prevajanju srbščina-slovenščina in obratno prihranek časa največji. To je za sorodna jezika pričakovano. Strojni prevodi so lahko zelo učinkovita vmesna faza pri prevajanju tudi za večino drugih jezikov.

Omenimo še en vidik uporabe strojnih prevodov. Za prevajalce popravljanje strojnih prevodov ni najbolj »všečen« proces in nekateri do tega čutijo določen odpor. V tem oziru so lahko prikazani rezultati do neke mere popačen prikaz, subjektivna percepcija strojnega prevajanja profesionalnih prevajalcev.

5. Zaključek

V članku smo predstavili sistem strojnega prevajanja za podnapise, ki smo ga razvili v projektu SUMAT. Podrobneje smo opisali evalvacijo kvalitete prevodov in učinkovitosti uporabe strojnih prevodov v prevajalskih procesih profesionalnih prevajalcev. Evalvacija s pomočjo prevajalcev je pokazala, da je kvaliteta prevodov za jezikovni par slovenščina-srbščina na visoki ravni. Povprečna ocena prevodov je več kot 3,5. Kar 40 % prevodov je dobilo oceno 4, kar pomeni, da je bilo potrebnih le malo popravkov za zagotavljanje običajne kvalitete prevodov.

Tudi produktivnost prevajalca se lahko z uporabo strojnih prevodov poveča, zahteva pa od prevajalca prilagajanje na nov način dela. Zaenkrat je popravljanje strojnih prevodov še relativno nepoznan postopek med prevajalci. Da bi bilo strojno prevajanje pozitivno sprejeto med njimi, bi bilo treba učenje tehnik popravljanja vključiti tudi v učne procese v prevajalstvu.

6. Literatura

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., 2007. Moses: Open source toolkit for statistical machine translation. *Zbornik 45th Annual Meeting of the ACL*, 177–180.

Och, F. J., 2003. Minimum error rate training in statistical machine translation, *Zbornik 41st Annual meeting of the Association for Computational Linguistics,* Sapporo, Japan.

Papineni, K., Roukos, S., Ward, T., Zhu. W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. *40th Annual meeting of the Association for Computational Linguistics*, Philadelphia, 311–318.

Sennrich, R., 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. *Zbornik 13th Conference of the European Chapter of the Association for Computational Linguistics,* 539–549.

Sepesy Maučec, M., Presker, M., Zimšek, D., Rojc, M., Vlaj, D., Verdonik, D., Kačič, Z., 2012. Izdelava slovensko-srbskega vzporednega korpusa podnapisov za razvoj strojnega prevajanja v projektu SUMAT. *V: ERJAVEC, T. (ur.), ŽGANEC GROS, J. (ur.). Zbornik Osme konference Jezikovne tehnologije*, 167–172.

Snover, M. G., Madnani, N., Dorr, B., in Schwartz, R., 2006: TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation* 23(2–3): 117–127.

Specia, L., Shah, K., de Souza, J. G., Cohn, T., Kessler, F. B., 2013. QuEst–a translation quality estimation framework. Zbornik 51st *Annual meeting of the Association for Computational Linguistics* : System Demonstrations, 79–84.

Stolcke, A., 2002. SRILM: an extensible language modeling toolkit. *Proceedings of the Int. Conf. on Spoken Language Processing*, 901–904.

Tiedemann, J., 2009. News from OPUS - a collection of multilingual parallel corpora with tools and interfaces, *Recent Advances in Natural Language Processing*, vol. V, 237–248.

Verdonik, D., Sepesy Maučec, M., 2013. O avtomatski evalvaciji strojnega prevajanja. *Slovenščina 2.0*, 2013, št. 1, 111–133.

Luščenje borzne terminologije

Senja Pollak*, Biljana Božinovski‡,

* Odsek za tehnologije znanja, Institut Jožef Stefan
Jamova 39, 1000 Ljubljana
senja.pollak@ijs.si
‡ Biljana Božinovski
Maistrova ulica 2, 8250 Brežice
bbozinovski@poslovniprevodi.si

Povzetek

V članku je predstavljen pristop h gradnji geslovnika za slovar borzne terminologije, izdelan na podlagi avtomatskega luščenja terminologije. Predstavimo korpus slovenskega borznega jezika ter motiviramo izbiro pristopa luščenja z orodjem LUIZ-CF, ki ga tudi primerjamo s pristopom, temelječim na orodju WordSmith Tools. Izračunamo natančnost in priklic avtomatskega luščenja ter strinjanje med ocenjevalcema. V nadaljevanju analiziramo izluščene termine ter podamo predloge za izboljšavo luščilnika.

Extraction of Stock Market Terminology

The paper presents an approach to building the wordlist for a dictionary of stock market terminology using automatic terminology extraction. A specialised corpus of Slovene stock market terminology is presented, followed by the argumentation why terminology was extracted with LUIZ-CF. In addition, the approach to building a wordlist using LUIZ-CF is compared with the approach using WordSmith Tools. Precision and recall are calculated for both term extraction methods, and the level of agreement between two evaluators is examined. We analyse the extracted terminology and propose improvements of the selected term extraction tool.

1. Uvod

Inventar terminologije strokovnega področja je osnova za izdelavo terminološkega slovarja oziroma terminološke zbirke, ki je ključni jezikovni vir strokovnega prevajalca. Izdelava geslovnika lahko poteka ročno, kar zahteva veliko strokovnega znanja, je časovno zamudno ter izhaja iz subjektivnih izbir. Zato so se v zadnjem desetletju raziskovalci s področja računalniškega jezikoslovja posvetili izdelavi metod za avtomatsko luščenje terminologije iz korpusov. Samodejne metode so bile razvite za različne jezike, npr. za angleščino Sclano in Velardi (2007), Ahmad idr. (2007), Frantzi in Ananiadou (1999), Kozakov idr. (2004); za slovenščino rešitve ponuja Vintar (2003, 2010). Dvo- in večjezične rešitve predstavljajo npr. Lefever idr. (2009), Macken idr. (2013), na voljo pa so tudi plačljiva orodja, kot so SDL MultiTerm Extract,[1] WordSmith Tools[2] in SketchEngine.[3]

Namen članka je opisati uporabo jezikovnih tehnologij, natančneje avtomatskega luščenja terminologije, pri izdelavi geslovnika slovenske borzne terminologije. Angleška borzna terminologija je dostopna v slovarjih (npr. Barron's Dictionary of Finance and Investment Terms) in zlasti na spletu, kjer svetovne borze (npr. ameriški NASDAQ, angleški LSE, kanadski TMX, avstralski ASX) in investicijski portali (Investopedia, Investor Words) predstavljajo tudi najnovejše strokovne izraze. Slovenska borzna terminologija je samostojno predstavljena v Borznih izrazih (Čas in Rotar, 1994); gre za večjezični slovar, ki je služil kot podlaga za vključitev borznih izrazov tudi v številne kasnejše spletne terminološke zbirke finančnih institucij in investicijskih portalov (NLB, Abanka, vzajemci.com in številni drugi). Omenjene zbirke, ki sicer pokrivajo širša področja financ, bančništva in podobno, imajo s terminološkega in terminografskega vidika nekaj pomanjkljivosti: vsebujejo borzne izraze, ki niso več v uporabi (npr. francoska tujka *fond*), ne vsebujejo številnih na novo skovanih izrazov, ki so se v slovenski borzni terminologiji ustalili zlasti od začetka finančne krize naprej (npr. *slaba banka*), ter niso izdelane v skladu z načeli terminografske stroke (zapisi terminov z velikimi začetnicami/tiskanimi črkami, ciklične/nepopolne/netočne definicije ipd.). Zaradi odsotnosti standarda in neenotne rabe se v besedilih pojavljajo številne dvojnice (*investitor/vlagatelj, borzna kotacija/uradna kotacija, tečaj/cena vrednostnega papirja*), ki nestrokovnjaka begajo in otežujejo prevajanje. Kaže se potreba po sistematični analizi sodobnih slovenskih besedil z borznega področja in zajemu aktualnega izrazja (ter drugih besedilnih informacij) v dejanski rabi, in sicer za nadgradnjo obstoječih oziroma izdelavo normativnega dvojezičnega slovarja, pomembnega terminološkega vira strokovnih prevajalcev. V članku opišemo začetno stopnjo izdelave slovenskega geslovnika dvojezičnega slovarja borznega jezika. Predstavimo dva pristopa za avtomatsko luščenje terminologije in ju na primeru luščenja iz specializiranega korpusa borznih besedil ovrednotimo.

2. Korpus borznega jezika

Korpus borznega jezika (Božinovski, 2014) je enojezični sinhroni, zaključeni, specializirani korpus. Vanj je vključenih 76 besedilnih dokumentov s področja trga kapitala v slovenskem jeziku, ki so nastala od leta 1999 dalje. Da bi korpusu, ki obsega 1.282.392 besed, zagotovili reprezentativnost oziroma uravnoteženost (Biber, 1993; Atkins idr., 1992; Arhar Holdt, 2006), smo besedila zajemali po vseh kategorijah tvorcev besedil s področja trga kapitala: zanimala so nas besedila profesorjev (ekonomistov, pravnikov) in študentov, besedila institucij trga kapitala (Ljubljanska borza, Agencija za trg vrednostnih papirjev, Centralna klirinško depotna družba, borzni člani, družbe za upravljanje,

[1] http://www.sdl.com/products/sdl-multiterm/extract.html
[2] WordSmith Tools (http://www.lexically.net/wordsmith/) v brezplačni različici omogoča omejen izpis rezultatov.
[3] SketchEngine (http://www.sketchengine.co.uk/) je v brezplačni različici na voljo 30 dni.

izdajatelji) in besedila zakonodajalca (zakonodaja) ter specializirana publicistična besedila (revija Kapital, časnik Finance itd.). Vpričo omejenih resursov smo se odločili v korpus vključiti vsa besedila, ki jih je bilo možno v razpoložljivem času pridobiti brezplačno in v primerni elektronski obliki. Izdelani korpus vsebuje znanstvene in strokovne monografije ter članke, študije in elaborate, srednje- in visokošolske učbenike, zaključna visokošolska dela, zakonodajo in predpise regulatorjev trga kapitala, brošure ter publicistična besedila. Držali smo se pravila, da se pri gradnji specializiranih korpusov zajemajo besedila v stroki uveljavljenih avtorjev (Atkins in Clear, 1992), in stremeli k čim bolj raznovrstnemu avtorstvu z namenom izključiti individualne posebnosti (Pearson, 1998). Korpus je v grobem razdeljen na tri podkorpuse: *znanstveni* vsebuje besedila uveljavljenih ekonomistov (37-odstotni delež celotnega korpusa), *strokovni* besedila borznih članov in izdajateljev ter zakonodajo in ostale predpise (42-odstotni delež), *poljudnostrokovni* pa publicistična besedila, brošure, spletne predstavitve in podobno (21-odstotni delež).[4] Večinoma so vključena celotna besedila, pri učbenikih in monografijah, ki pokrivajo področja, širša od kapitalskih trgov, pa smo poglavja, ki obravnavajo druge teme, izpustili. Vse dokumente, vključene v korpus, smo pridobili v elektronski obliki in jih tudi s pomočjo optičnih čitalnikov pretvorili v končno golo besedilo, kodirano v utf-8. Ker za vsa besedila nismo pridobili dovoljenja za javno objavo, je korpus interne narave.

3. Izbira pristopa za luščenje terminologije

K izdelavi osnutka geslovnika slovarja borznega jezika smo pristopili z avtomatskim luščenjem terminologije iz omenjenega korpusa. Po pregledu orodij za luščenje terminologije, ki podpirajo slovenščino, smo izbrali dve metodi: luščenje z orodjem WordSmith Tools, ki ga uporablja Sekcija za terminološke slovarje na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU (v nadaljevanju STS SAZU), ter orodje LUIZ, ki ga v implementaciji znotraj delotoka v okolju ClowdFlows imenujemo LUIZ-CF.

3.1. WordSmith Tools (WS-L)

WordSmith Tools (različica 6, Scott, 2014) je napreden program za analizo korpusnih podatkov. Kot vstopno točko pri delu s terminologijo v specializiranem korpusu ga omenja in uporablja več avtorjev (Vintar, 2008: 92, Snoj, 2013: 2, 4, 6). Orodje smo uporabili skladno s prakso STS SAZU:[5]

V program smo uvozili celoten korpus kot golo besedilo (utf-8) ter preizkusili komponento Wordlist, ki na podlagi korpusa besedil sestavi seznam besed, urejenih po pogostosti. Gre za osnovno informacijo, ki jo potrebujemo v začetni fazi ukvarjanja s terminologijo v specializiranem korpusu (Vintar, 2008: 92). Korpus smo lematizirali z uvozom seznama lem[6] in pripadajočih besednih oblik, poleg lematizacije pa Wordlist omogoča tudi uvoz seznama "praznih besed", oz. v našem primeru seznam splošnih besed,[7] saj za razliko od običajnih seznamov praznih besed, ki vsebujejo predvsem veznike, členke, predloge itd., seznam, ki smo ga uvozili mi, zajema mnogo obširnejši nabor besed, vključujoč poleg t.i. praznih besed tudi številna lastna imena, pridevnike in druge besede splošnega jezika, s čimer se seznam bolj prilagodi terminološki nalogi.[8] Wordlist nam omogoča le luščenje enobesednih terminov,[9] v nadaljevanju pa omenjeni pristop okrajšamo z WS-L.

3.2. LUIZ-CF

Drugo orodje, ki smo ga uporabili, je LUIZ-CF, kakor imenujemo reimplementacijo sistema LUIZ (Vintar, 2010) v obliki prosto dostopnega delotoka v okolju ClowdFlows (Kranjc idr., 2012). Orodje LUIZ-CF je prosto dostopno[10] v okviru luščilnika terminologije in definicij (Pollak idr. 2012a, Pollak 2014).

Luščilnik na podlagi oblikoskladenjskih vzorcev izdela nabor kandidatov za termine, ki jih nato razvrsti glede na izračun njihove terminološke vrednosti, pri čemer primerja njihovo frekvenco v danem (specializiranem) in referenčnem korpusu[11] (Vintar, 2010). Za namen raziskave smo uporabili le del delotoka, potreben za luščenje terminologije (brez gradnikov za luščenje definicij): preko Load Corpus smo korpus naložili kot golo besedilo (utf-8), v gradniku ToTrTaLe[12] za jezikovno označevanje ter v zadnjem, ključnem gradniku Term Extraction pa smo izbrali jezik slovenščina. Kot rezultat sistema LUIZ-CF uporabnik dobi seznam terminoloških kandidatov, na katerem so eno- in večbesedni terminološki kandidati, razvrščeni na istem seznamu z

[4] Poimenovanja *znanstveni*, *strokovni* in *poljudnostrokovni* nimajo posebne metodološke vrednosti, uporabljamo jih za interno kategorizacijo besedil.

[5] Avtorica B. B. sem pristop zasnovala in izvedla skladno z opisom metodologije STS SAZU, podanim s strani dr. Tanje Fajfar, raziskovalke v STS SAZU, v osebni korespondenci dne 23. 4. in 15. 9. 2014.

[6] Uporabljeni seznam lem vsebuje 842.091 besednih oblik (100.784 lem) in je bil izdelan na podlagi leksikona besednih oblik za slovenski jezik Sloleks: http://www.slovenscina.eu/sloleks. Ker uporabljeni seznam ne vsebuje strokovnih izrazov našega področja (npr. ID, SEOnet, ETF, certifikat), opisana lematizacija pri teh kandidatih ni delovala in smo jo naknadno deloma opravili ročno.

[7] Uporabljeni seznam, ki smo ga vnesli kot seznam "praznih besed", vsebuje 343.963 besednih oblik in je bil izdelan s pomočjo Slovarja slovenskega knjižnega jezika in Slovenskega pravopisa. Oba seznama nam je zagotovil dr. A. Perdih z Inštituta za slovenski jezik Frana Ramovša ZRC SAZU.

[8] WordSmith Tools poleg osnovne funkcije Wordlist vsebuje še možnost iskanja večbesednih skupkov Clusters ter komponento Keywords, ki identificira besede v korpusu po ključnosti. Slednjih po našem vedenju na STS SAZU ne uporabljajo, zato se tudi sami osredotočimo na funkcijo Wordlist, drugi dve funkciji pa le na kratko preizkusimo. Funkcijo Keywords uporabljeni pristop delno nadomesti z zgoraj opisanim obsežnim seznamom splošnih besed. V nadaljnjem delu pa bomo v podrobnejšo primerjavo vključili tudi drugi komponenti.

[9] Osnutki geslovnikov, ki nastajajo v okviru STS SAZU, sicer vsebujejo večbesedne termine, ki jih na podlagi Wordlista preko funkcije Concord terminologi dodajo ročno.

[10] http://www.clowdflows.org/workflow/1380/

[11] V trenutni implementaciji se uporablja referenčni korpus FidaPLUS (Arhar Holdt in Gorjanc, 2007).

[12] Gradnik implementira orodje ToTrTaLe (Erjavec, 2011) in je podrobneje opisan v Pollak idr. (2012b). Izbrali smo tudi parameter za post-procesiranje, kot izhodni format pa ".txt".

normalizirano terminološko vrednostjo med 1 (najboljši) in 0 (najslabši kandidat).

3.3. Primerjava pristopov

Pristopa sta z merama natančnosti (angl. precision) in priklica (angl. recall) ovrednotena v Tabelah 1 in 2.[13] Za izračun ocene natančnosti obeh pristopov za namen sestave geslovnika smo izluščenim terminološkim kandidatom pripisali vrednosti 1 (termin) ali 0 (ni termin). V geslovnik bodo vključeni borzni termini (prim. ime stolpca *Borzni* v Tabeli 1), ovrednotili pa smo tudi termine s področij, sorodnih borznemu (korporativno pravo, računovodstvo, finance), ki sicer verjetno ne bodo vključeni v geslovnik slovarja borzne terminologije, a so z vidika terminologije vseeno zanimivi (stolpec *Vsi* tako označuje odstotek izluščenih borznih in sorodnih terminov).

Najboljši kandidati so ponavadi tisti pri vrhu seznamov. Težje pa je zajeti manj pogoste termine, zato smo za izračun ocene natančnosti pri obeh pristopih ocenili 600 kandidatov iz različnih delov seznama (imena vrstic v Tabeli 1 kažejo na zaporedno številko prvega od stotih ocenjenih kandidatov).[14]

Za izračun ocene priklica smo v naključnem dokumentu[15] iz korpusa ročno označili vse borzne termine. Tabela 2 prikaže, kolikšen odstotek tako označenih kandidatov izluščimo s posameznim orodjem ter na katerih nivojih seznamov jih najdemo (med vrhnjimi 1000, med vrhnjimi 2000 kandidati itd.).

OCENA NATANČNOSTI	WS-L		LUIZ-CF	
	Borzni	**Vsi**	**Borzni**	**Vsi**
Nivo 1	0,36	0,47	0,56	0,67
Nivo 1000	0,12	0,18	0,44	0,66
Nivo 2000	0,12	0,25	0,25	0,46
Nivo 3000	0,10	0,19	0,19	0,40
Nivo 4000	0,05	0,20	0,19	0,41
Nivo 5000	0,04	0,10	0,21	0,31
Vsi (600 kandidatov)	**0,13**	**0,23**	**0,31**	**0,48**

Tabela 1: Rezultati natančnosti luščenja (v odstotkih).

OCENA PRIKLICA	Vsi		Enobesedni	
	WS-L	**LUIZ-CF**	**WS-L**	**LUIZ-CF enobesedni**
Vrhnjih 100	0,11	0,24	0,45	0,65
Vrhnjih 1000	0,20	0,51	0,85	0,90
Vrhnjih 2000	0,23	0,70	0,95	1,00
Vrhnjih 3000		0,72		
Vrhnjih 4000		0,75		
Vrhnjih 5000		0,78		
Vrhnjih 10000		0,85		

Tabela 2: Rezultati priklica luščenja (v odstotkih).

Kot kaže Tabela 1, je pristop z orodjem LUIZ-CF bolj natančen od pristopa z orodjem Wordlist. Med vrhnjimi 100 terminološkimi kandidati je orodje LUIZ-CF namreč izluščilo več kot polovico (56 odstotkov) borznih terminov, kar je 20 odstotkov več kot pri pristopu WS-L. Tudi na nižjih nivojih je več terminov na seznamu LUIZ-CF, tako borznih kot sorodnih. Kljub temu, da natančnost na nižjih mestih večinoma pada, je med kandidati, izluščenimi z orodjem LUIZ-CF, še zmeraj približno petina borznih terminov oziroma tretjina, če upoštevamo tudi termine sorodnih področij.

Pristopa smo v nadaljevanju primerjali tudi glede priklica na podlagi ročno označenega dokumenta. Ugotovili smo, da če upoštevamo tako enobesedne kot večbesedne termine (skupaj 83 terminov), LUIZ-CF izlušči bistveno večji odstotek terminov kot WS-L. Med vrhnjimi 1000 kandidati najdemo dobro polovico terminov iz ročno označenega dokumenta (orodje WS-L izlušči le 20 odstotkov označenih terminov). Med vsemi izluščenimi kandidati, torej vključno s tistimi z zelo nizko terminološko vrednostjo, tj. do mesta 10000, je 85 odstotkov preverjanih terminov (za manjkajoče je v veliki meri kriv nepopoln nabor oblikoskladenjskih vzorcev, ki jih luščilnik zajema). Velika razlika med orodjema izhaja predvsem iz luščenja zgolj enobesednih kandidatov z orodjem WS-L v nasprotju z luščenjem tako eno- kot večbesednih z orodjem LUIZ-CF. Zaradi doslednosti primerjave in jasne utemeljitve izbire orodja za gradnjo geslovnika smo posebej izračunali še priklic za le enobesedne termine. Desni del Tabele 2 (*Enobesedni*) torej prikazuje delež izluščenih enobesednih terminov od vseh enobesednih terminov v izbranem testnem dokumentu (20 terminov). WS-L seznam tako in tako vključuje le enobesedne termine, seznam orodja LUIZ-CF pa smo skrčili na seznam enobesednih kandidatov za namen te primerjave. Na podlagi Tabele 2 smo se tudi z vidika priklica odločili za nadaljevanje sestave geslovnika z orodjem LUIZ-CF.

Za vrhnjih 100 terminoloških kandidatov vsakega orodja smo izračunali tudi stopnjo ujemanja med dvema ocenjevalcema. Poleg polstrokovnjaka prevajalca je seznam ovrednotil še področni strokovnjak ekonomist. Rezultati kažejo, da je strokovnjak kot termine označil manj kandidatov kot polstrokovnjak prevajalec, kar je bilo pričakovano in kar potrjujejo tudi izkušnje drugih (prim. Vintar, 2003; Logar Berginc idr., 2013). Kvantitativna razlika med pristopoma se je potrdila tudi pri tej oceni: na seznamu LUIZ-CF je strokovnjak kot borzne termine označil 23 odstotkov kandidatov, kot borzne in sorodne termine skupaj pa 53 odstotkov kandidatov (prim. z odstotkoma 0,56 in 0,67 pri polstrokovnjaku v Tabeli 1), medtem ko je bil pristop WS-L ocenjen bistveno slabše: potrjenih terminov je bilo 26, od tega borznih le 10 odstotkov seznama vrhnjih 100 kandidatov. Na seznamu, ki smo ga zgradili iz različnic zgornjih 100 kandidatov vsakega orodja (skupaj 140 terminov), smo izračunali splošno strinjanje (mera, ki v odstotkih izraža primere, ko sta oba ocenjevalca kandidat označila kot termin oz. netermin) in dobili rezultat 0,77. Zanimal nas je tudi koeficient kappa (Cohen, 1960), ki upošteva razliko med naključno verjetnostjo strinjanja ter opaženim strinjanjem. Za izračun kappe smo uporabili spletno orodje Vassarstats (Lowry, 2013). Kappa 0 pomeni naključno strinjanje, 1 pa popolno strinjanje. Naš rezultat je 0,5 in izraža srednjo stopnjo strinjanja (angl. moderate agreement, glej Viera in Garrett (2005)).

Da bi se dokončno prepričali o pravi izbiri orodja za nadaljevanje dela, smo na hitro preizkusili tudi drugi

[13] Ena izmed ocenjevalk v eksperimentih je strokovna prevajalka B. B., soavtorica pričujočega članka in avtorica nastajajočega borznega slovarja.

[14] Evalvacija je obsegala 600 izluščenih kandidatov na zaporednih mestih 1–100, 1000–1099, 2000–2099, 3000–3099, 4000–4099 in 5000–5099 obeh generiranih seznamov.

[15] Ljubljanska borza, d. d.: Vodnik za vlagatelje na Ljubljanski borzi, http://www.ljse.si/cgi-bin/jve.cgi?att=16774. V dokumentu z 2000 besedami je B. B. identificirala 83 terminov.

komponenti orodja WordSmith Tools, namreč funkciji Clusters in Keywords. Rezultate smo ovrednotili na naboru vrhnjih 1000 kandidatov. Funkcija za iskanje večbesednih skupkov Clusters je delovala bistveno slabše, komponenta Keywords za sezname ključnih besed pa sicer nekoliko boje od preizkušene funkcije Wordlist, vendar še vedno bistveno slabše od LUIZ-CF.

Seveda bi bilo za metodološko koherentno primerjavo samih orodij potrebno ločiti luščenje od drugih korakov (uporabiti enake načine lematizacije korpusa, upoštevati vpliv seznamov praznih besed in referenčnih korpusov itd.). Toda osnovni namen naše primerjave je bil izbrati pristop za gradnjo geslovnika strokovnega slovarja, zato smo se omejili le na primerjavo pristopov, ki so nam bili na voljo brez dodatnega dela. Na podlagi pridobljenih rezultatov smo izbrali pristop z uporabo orodja LUIZ-CF, ki tudi ne zahteva izdelave nobenih dodatnih seznamov ali predprocesiranja in je preprosto za uporabo.

4. Izdelava geslovnika in interpretacija rezultatov

S pristopom za luščenje terminologije na podlagi luščilnika LUIZ-CF smo izluščili dobrih 11000 kandidatov, od katerih smo jih ročno pregledali vrhnjih 6000. Prvih 25 terminoloških kandidatov je prikazanih v Tabeli 3. V nadaljevanju navajamo nekatera spoznanja, oblikovana med ročnim pregledovanjem izluščenih kandidatov.

Ena prvih dilem, s katero smo soočeni med ročnim pregledovanjem, je, kje in kako postaviti mejo med terminološko in splošno leksiko ter med borzno in sorodno terminologijo. Terminološka kandidata, ki ju je LUIZ-CF uvrstil na sam vrh seznama (oba imata oceno 1.0) sta *družba* in *vrednostni papir*. Medtem ko drugi ni sporen, saj gre za izrazito borzni termin, si prvega deli več strok, sega pa tudi na območje splošnega jezika; v korpusu borznega jezika so izpričani tako 1. in 2. pomen, naveden v SSKJ, ter več pravnih opredelitev družbe. Ker sestavljamo slovar borznega jezika, nas ne zanima nujno celota pomenov, zajetih v korpusu, temveč le pomeni, vezani na ožje borzno področje. Tako se bodo v geslovniku (po posvetovanju s področnim strokovnjakom) predvidoma znašli termini *borzna družba, borznoposredniška družba, delniška družba, družba za upravljanje, javna družba, investicijska družba, kapitalska družba* itd., ne pa tudi *ciljna družba, holdinška družba, hčerinska družba, družba z omejeno odgovornostjo* itd., ki spadajo na sorodno področje korporativnega prava, in to četudi imajo nekateri kandidati iz druge skupine morda višjo terminološko vrednost. Meja med termini sorodnih področji in borznimi termini je pogosto zelo tanka, saj se področje kapitalskih trgov nahaja na presečišču več strokovnih področij (prim. Logar Berginc idr., 2013, ki podobno ugotavljajo za področje odnosov z javnostmi). Po ocenah prve ocenjevalke in kot je razvidno iz zadnje vrstice Tabele 1, jih med 600 ocenjenimi kandidati za termine, ki jih je izluščilo orodje LUIZ-CF, 48 odstotkov predstavlja relevantno terminologijo: 31 odstotkov je borznih terminov in 17 odstotkov terminov s sorodnih področij (npr. *dolžniška kriza, prevzemna ponudba, bilanca stanja*), pri čemer imajo termini sorodnih področij lahko zelo visoke terminološke vrednosti. Za prepoznavanje mej med terminologijo različnih področij je potrebno termine opazovati v sobesedilu ter upoštevati mnenje stroke.

Term. vrednost	Lema	Kanonična oblika
1.000000	[družba]	<<družba>>
1.000000	[vrednosten papir]	<<vrednostni papir>>
0.671458	[trg]	<<trg>>
0.581797	[delnica]	<<delnica>>
0.399937	[člen]	<<člen>>
0.253923	[papir]	<<papir>>
0.206269	[sklad]	<<sklad>>
0.169046	[banka]	<<banka>>
0.156982	[borza]	<<borza>>
0.151592	[kapital]	<<kapital>>
0.143811	[podjetje]	<<podjetje>>
0.140058	[finančen instrument]	<<finančni instrument>>
0.127859	[nadzoren svet]	<<nadzorni svet>>
0.119758	[obveznica]	<<obveznica>>
0.109289	[odstavek]	<<odstavek>>
0.081915	[instrument]	<<instrument>>
0.070904	[trgovanje]	<<trgovanje>>
0.066481	[vrednost]	<<vrednost>>
0.066472	[tveganje]	<<tveganje>>
0.063684	[delničar]	<<delničar>>
0.057922	[član]	<<član>>
0.054564	[posel]	<<posel>>
0.054328	[naložba]	<<naložba>>
0.053878	[vlagatelj]	<<vlagatelj>>
0.052707	[obresten mera]	<<obrestna mera>>

Tabela 3: Vrhnjih 25 kandidatov, izluščenih z LUIZ-CF.

Med ročnim pregledovanjem izluščenih kandidatov smo obdržali le termine z ožjega področja kapitalskih trgov in dodali manjkajoče izraze z namenom dopolnitve posameznih pojmovnih polj ter druge ključne izraze področja. Če bi želeli v celoti slediti korpusnemu pristopu, bi morali korpus ciljno dopolnjevati z besedili, ki vsebujejo manjkajoče izraze, česar nam časovni okvir ne dopušča, poleg tega je to postopek, ki se verjetno nikoli ne konča (Atkins in Rundell 2008).

Trenutno poteka druga faza usklajevanja izrazov s področnim strokovnjakom; skupno število že potrjenih terminov za geslovnik je 1262, od katerih smo jih 91 odstotkov izluščili s pomočjo LUIZ-CF, 9 odstotkov pa jih je bilo vključenih naknadno.

Značilnost borzne terminologije, ki smo jo opazili pri ročni izdelavi geslovnika, je soobstoj številnih terminoloških variacij (npr. *prvi trgovalni dan brez upravičenja do dividend/datum brez dividend/eksdividendni datum*), kar kaže na dolgoletno odsotnost standarda. Glede rabe so korpusni podatki v nekaterih primerih v popolnem nasprotju s prepričanjem stroke (stroka denimo zadnja leta zagovarja rabo termina *finančni instrument* namesto *vrednostni papir*, četudi je slednji bistveno pogostejši v praktično vseh večbesednih zvezah, npr. *dolžniški vrednostni papir* se v korpusu pojavi 247-krat, *dolžniški finančni instrument* pa 12-krat). Če želi nastajajoči borzni slovar prevzeti normativno vlogo, bo zato potrebno soočiti rabo, kot se je z leti ustalila, in mnenje stroke.

Seznam izluščenih terminoloških kandidatov nam je sprva služil tudi kot osnova za dopolnitev terminološke

zbirke s terminološkimi kolokacijami.[16] Vendar se je upoštevanje kolokacij na začetni stopnji sestave geslovnika izkazalo za preveč zamudno, saj je zaradi velike količine pojmovno nerazvrščenih kandidatov oteževalo delo. Zato smo jih na tej stopnji izključili iz obravnave in se bomo k njim vrnili v fazi izdelave geselskih člankov.

Podkorpus PS	Podkorpus S	Podkorpus Z
delnica	družba	trg
milijon evrov	vrednostni papir	vrednostni papir
vrednostni papir	član	podjetje
obrestna mera	nadzorni svet	kapital
evro	finančni instrument	delnica
trg	odstavek	banka
odstotek	delnica	papir
milijarda evrov	borznoposredniška družba	obrestna mera
ljubljanska borza	papir	obveznica
obveznica	član	naložba

Tabela 4: Vrhnjih 10 kandidatov po podkorpusih (poljudnostrokovni, strokovni, znanstveni).

Dodatno se nam je zdelo zanimivo preučiti rezultate luščenja terminologije za posamezne podkorpuse, torej za podkorpus znanstvenih besedil, podkorpus strokovnih besedil in podkorpus poljudnostrokovnih besedil. Primerjali smo vrhnjih 100 izluščenih kandidatov vsakega podkorpusa, po prvih 10 predstavljamo v Tabeli 4. Analiza terminov kaže, da so besedila po strukturi besedišča v posameznih podkorpusih zelo raznolika. Največja opazna odstopanja so pravno-zakonodajni izrazi (*člen, odstavek, določba, zakon*) v podkorpusu strokovnih besedil, v katerem prevladujejo predpisi s področja trga kapitala, ter izrazi za vrednosti (*milijon/milijarda evrov/dolarjev*) v podkorpusu poljudnostrokovnih besedil. Največji delež pravilno izluščenih terminov je sicer luščilnik prepoznal v znanstvenem podkorpusu (79 %).

5. Predlogi za izboljšavo luščilnika

V tem razdelku analiziramo pomanjkljivosti luščenja terminologije z LUIZ-CF z vidika uporabnosti za sestavo geslovnika strokovnega področja.

Obravnava terminoloških variacij. Za borzno terminologijo je v odsotnosti standarda značilna variantnost, saj raba ni ustaljena. LUIZ-CF variacije prikaže kot različne kandidate, ki jih je treba ročno iskati in združevati. Gre tako za pisne variacije (npr. finančna institucija vs. finančna inštitucija, SEOnet vs. SEO-net vs. seo-net vs. SEO net) kot oblikoslovne (posoja/posojanje vrednostnih papirjev) in skladenjske variacije (kapitalski trg vs. trg kapitala).

S tem je tesno povezana *obravnava enakovrednih poimenovanj.* V izluščeni terminologiji je mnogo primerov, ko za en pojem obstajata dve različni enakovredni poimenovanji, ki jih LUIZ-CF prikaže neodvisno drugo od drugega. Gre za pare, za katere bi bilo z vidika pojmovnega pristopa zaželeno, da bi jih orodje ponudilo skupaj. Glavni podkategoriji sta domač izraz – tujka (npr. vlagatelj/investitor, izračun/kliring, navzkrižje/konflikt interesov) in razvezan izraz – kratica (celotna globina trga/CGT, vzajemni sklad/VS).

Lastna imena LUIZ-CF enakovredno razvršča med kandidate za termine, vključno z osebnimi imeni (priimki: Berk, Simoneti; imena indeksov: Eurostoxx, Dow Jones, NASDAQ) in imeni institucij (Ljubljanska borza, Wall Street, Ameriška centralna banka), pri katerih je odločitev za ali proti vključitvi v slovar zahtevna. Koristen bi bil parameter, s katerim bi lahko seznama ločili.

Sorodni termini. Kot že večkrat omenjeno zgoraj, je težava tovrstnega luščenja tudi nehoten zajem številnih kandidatov s sorodnih področij (nekaj primerov parov borznih/neborznih terminov: depozitar/depozit, trgovalni račun/tekoči račun, delniška družba/komanditna družba, devizna izmenjava/devizni tečaj). Pri nadgradnji LUIZ-CF bi bilo v ta namen smiselno preizkusili strojne metode aktivnega učenja (angl. active learning).

Terminološke kolokacije. V fazi izdelave geslovnika so problematične tudi terminološke kolokacije, saj se izkaže, da na začetni stopnji gradnje slovarja zmanjšajo preglednost zbranega gradiva in upočasnjujejo inventarizacijo terminologije. LUIZ-CF jih je izluščil zelo veliko (primeri parov borzni termin/terminološka kolokacija: osnovni kapital/povečanje osnovnega kapitala, skupščina delničarjev/sklic skupščine delničarjev, avkcija/trg v avkciji). V kolikor pride do odločitve za vključitev kolokacij v terminološki slovar, so te potrebne šele kasneje in ne že v fazi sestave geslovnika, zato bi bilo smiselno, da so na seznamu izluščenih kandidatov za termine prikazane ločeno.

Omenili smo že, da priklic ni bil 100 odstoten zaradi nekaterih nepokritih *oblikoskladenjskih vzorcev*.

6. Zaključki in nadaljnje delo

V članku smo predstavili poskus luščenja terminologije iz korpusa borznega jezika, ki ga sestavljajo besedila znanstvenega, strokovnega in poljudnostrokovnega značaja. Terminologijo smo luščili z uporabo luščilnika LUIZ (Vintar, 2010) v spletni implementaciji LUIZ-CF (Pollak idr. 2012a, Pollak 2014), ter njegovo natančnost in priklic primerjali s podobnim pristopom z orodjem Wordlist (WordSmith Tools). Za temeljito primerjavo orodij bi morali pri obeh orodjih ločiti luščenje od predprocesiranja ter v skupku orodij WordSmith Tools natančneje preizkusiti še funkciji Keywords in Clusters.

Na podlagi rezultatov smo za izdelavo nastajajočega geslovnika slovarja borznega jezika izbrali orodje LUIZ-CF, terminologijo pa v sodelovanju s področnim strokovnjakom dopolnjujemo ročno. Analizirali smo lastnosti izluščenih terminoloških kandidatov in identificirali mesta za možne izboljšave luščilnika (obravnava terminoloških variacij, terminov sorodnih področij, kolokacij). Vzporedno izboljšujemo tudi sam delotok, da bi luščenje terminologije potekalo hitreje. V

[16] Izbirali smo jih izrazito selektivno, saj je namen njihove vključitve točno določen in dvojen, prilagojen zlasti uporabniku nestrokovnjaku: po eni strani naj bi z dodatnimi informacijami o pojmovnem polju termina okrepile razumevanje njegove razlage (Bergenholtz in Tarp, 1995) in uporabniku ponudile leksikalno okolje termina (Atkins in Rundell, 2008), po drugi strani pa podprle prevajalsko funkcijo nastajajočega slovarja (za ta namen bodo izbrane kontrastivno zanimive, torej netransparentne kolokacije). V zbirko kljub njihovi terminološkosti nismo želeli vključiti vseh kolokacij s področja kapitalskih trgov, temveč le tiste, ki podpirajo omenjena cilja.

nadaljevanju raziskave bomo delotok, predstavljen v Pollak (2014), uporabili za luščenje definicij iz omenjenega korpusa ter ovrednotili njegovo uporabnost.

Zahvala

Zahvaljujemo se dr. Tanji Fajfar, dr. Mojci Žagar Karer in dr. Andreju Perdrihu z Inštituta za slovenski jezik ZRC SAZU, ki so drugi avtorici članka prijazno pomagali in ji posredovali informacije o svojem pristopu k uporabi orodja WordSmith Tools ter ji omogočili uporabo svojih seznamov lem in praznih besed.

7. Literatura

Ahmad, K., L. Gillam, in L. Tostevin, 2007. University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Rerieval (WILDER). *Proceedings of the Eight Text REtrieval Conference (TREC-8)*: 717–724.

Arhar Holdt, Š., 2006. Gradnja specializiranega korpusa. *Jezik in slovstvo,* 51(1): 53–67.

Arhar Holdt, Š., in V. Gorjanc, 2007. Korpus FidaPLUS: Nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52 (2): 95–110

Atkins, S., in J. Clear, 1992. Corpus design criteria. *Literary and Linguistic Computing*, 7(1): 1–16.

Atkins, B.T.S, J. Clear, in N. Ostler, 1992. Corpus design criteria. *Literary and Linguistic Computing. Journal of the Association for Literary and Linguistic Computing* 7(1): 1–16.

Atkins, B.T.S, in M. Rundell, 2008. *The Oxford guide to practical lexicography*. Oxford/New York: Oxford University Press.

Bergenholtz, H., in S. Tarp (ur.), 1995. *Manual of Specialised Lexicography*. Amsterdam/Philadelphia: Benjamins Publishing Company.

Biber, D., 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing,* 8(4): 243–257.

Božinovski, B. 2014 (v pripravi). *Problematika slovensko-angleškega strokovnega izrazoslovja s področja borznega poslovanja*. Doktorska disertacija, Univerza v Ljubljani.

Cohen, J., 1960. A Coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20: 37–46.

Čas, M., in T. Rotar, 1994. *Borzni izrazi: s trojezičnim slovarjem*. Maribor: Kapital.

Erjavec, T., 2011. Automatic linguistic annotation of historical language: ToTrTaLe and XIX century Slovene. *Proceedings of the ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (ACL 2011)*.

Frantzi, K. T., in S. Ananiadou, 1999. The CValue/NCValue domain independent method for multi-word term extraction. *Journal of Natural Language Processing*, 6(3): 145–179.

Kozakov, L., Y. Park, T. Fin, Y. Drissi, Y. Doganata, in T. Cofino, 2004. Glossary extraction and utilization in the information search and delivery system for IBM technical support. *IBM Systems Journal*, 43(3): 546–563.

Kranjc, J., V. Podpečan, in N. Lavrač, 2012. ClowdFlows: A cloud based scientific workflow platform. *Proceedings of ECML/PKDD-2012 (2)*, Springer LNCS 7524: 816–819.

Lefever, E., M. Lieve, in V. Hoste, 2009. Language-independent bilingual terminology extraction from a multilingual parallel corpus. *Proceedings of the 12th Conference of the European Chapter of the ACL*: 496–504.

Logar Berginc, N., Š. Vintar, in Š. Arhar Holdt, 2013. Terminologija odnosov z javnostmi: korpus - luščenje - terminološka podatkovna zbirka. *Slovenščina 2.0*, 1(2): 113–138.

Lowry, R.. 2013. *Kappa as a measure of concordance in categorical sorting.* http://vassarstats.net/kappa.html. Zadnji dostop: 19. september, 2014.

Macken, L., E. Lefever, in V. Hoste, 2013. TExSIS: Bilingual terminology extraction from parallel corpora using chunk-based alignment. *Terminology,* 19(1): 1–30.

Pearson, J., 1998. *Terms in Context*. Amsterdam/Philadelphia: John Benjamins.

Pollak, S., A. Vavpetič, J. Kranjc, N. Lavrač, in Š. Vintar, 2012a. NLP workflow for online definition extraction from English and Slovene text corpora. *Proceedings of the 11th Conference on Natural Language Processing*, 53–60.

Pollak, S., N. Trdin, A. Vavpetič, in T. Erjavec, 2012b. NLP Web Services for Slovene and English: Morphosyntactic tagging, lemmatisation and definition extraction. *Informatica,* 36: 441–449.

Pollak, S., 2014. *Polavtomatsko modeliranje področnega znanja iz večjezičnih korpusov.* Doktorska disertacija, Univerza v Ljubljani.

Sclano, F., in P. Velardi, 2007. TermExtractor: a Web application to learn the common terminology of interest groups and research communities. *Proceedings of the 9th Conference on Terminology and Artificial Intelligence TIA 2007*: 8–9.

Scott, M., 2014a. *WordSmith Tools version 6* (and *WordSmith Tools Manual*), Liverpool: Lexical Analysis Software.

Snoj, M., 2013. Zaključno poročilo raziskovalnega projekta – 2013. Oznaka poročila: ARRS-RPROJ-ZP-2013/203.

Viera, A. J., in J. M. Garrett, 2005. Understanding interobserver agreement: The Kappa Statistic. *Family Medicine*, 37(5): 360–363.

Vintar, Š., 2003. *Uporaba vzporednih korpusov za računalniško podprto ustvarjanje dvojezičnih terminoloških virov*. Doktorska disertacija, Univerza v Ljubljani.

Vintar, Š., 2008. *Terminologija: terminološka veda in računalniško podprta terminografija*. Ljubljana: Znanstvena založba Filozofske fakultete.

Vintar, Š., 2010. Bilingual term recognition revisited. The bag-of-equivalents term alignment approach and its evaluation. *Terminology*, 16(2): 141–158.

Analiza uporabe slovničnih pregledovalnikov za slovenščino

Mario Jurišić,† Špela Vintar*

†Zariška ulica 17, 4000 Kranj
mario_jurisic10@hotmail.com
*Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
spela.vintar@ff.uni-lj.si

Povzetek

V prispevku smo analizirali uporabo obstoječih slovničnih pregledovalnikov za slovenščino, Besane in LanguageToola, z vidika natančnosti, zanesljivosti in praktične uporabnosti. Prav tako smo preverili hipotezo, da uporabniki danih orodij pogosteje upoštevajo predloge popravkov omenjenih orodij pri prevajanju v tuji jezik kot pri prevajanju v materni jezik. Prispevek ponuja možne odgovore na vprašanja o učinkovitosti in smiselnosti uporabe tovrstnih orodij, ki temeljijo na izsledkih praktičnega poskusa, v okviru katerega smo pričujoča orodja z vidika omenjenih parametrov proučevali pri prevajanju iz slovenščine v angleščino in iz angleščine v slovenščino. Izsledki danega poskusa upravičujejo uporabo obeh slovničnih pregledovalnikov za slovenščino, vendar obenem opozarjajo na posamezne možnosti za izboljšave teh orodij.

Analyzing the Use of Grammar Checkers for Slovenian

This paper presents an analysis of the existing grammar checkers for Slovenian, namely Besana and LanguageTool, in terms of their precision, reliability, and practical usefulness. The hypothesis was assessed claiming that the users of these tools trust the suggested error corrections more often when translating into a foreign language compared to translating into their mother tongue. The paper provides plausible answers to the questions on the efficiency and the point of using such tools based on the findings of a practical experiment. In the experiment, these tools were examined from the point of view of the listed parameters when translating from Slovenian into English and from English into Slovenian. The results justify the use of both grammar checkers for Slovenian, but they also suggest some possible upgrades to the tools discussed.

1. Uvod

V zadnjem času smo priča hitremu razvoju tako raziskovalno kot tudi tržno zanimivih jezikovnih tehnologij, kot so na primer strojno prevajanje, razpoznavanje in sinteza govora ter mnoge druge, med katere spada tudi slovnično pregledovanje besedil. Številne med njimi so med seboj tesno povezane, saj se njihovi temeljni sestavni deli v veliki meri prekrivajo. Strojni prevajalniki denimo za dodeljevanje in izbiranje najverjetnejših oblikoskladenjskih oznak besed uporabljajo označevalnike in stavčne analizatorje, na podlagi katerih so zgrajeni tudi slovnični pregledovalniki (Holozan, 2013). Izboljšave slovničnih pregledovalnikov so zato precej odvisne tudi od razvoja drugih tehnologij za računalniško obdelavo jezika.

Preverjanje slovnične ustreznosti besedil zajema prepoznavanje primerov napačne rabe jezika in – do določene mere – tudi predlaganje popravkov pri ustrezno zaznanih napakah (Grammar Checker - Definition and Examples of Grammar Checkers; Helfrich in Music, 2000). Ob tem je treba poudariti, da so slovnični pregledovalniki omejeni na odkrivanje nepravilnosti na pravopisni in skladenjski ravni jezika, kajti ta orodja trenutno še niso zmožna zaznati pomenskih napak, ki segajo preko meja povedi (Burston, 1996: 106).

Kljub temu je mogoče trditi, da slovnični pregledovalniki verjetno predstavljajo skupino najpogosteje uporabljanih pripomočkov za pisanje besedil. To ne velja samo v prevajalski panogi, temveč tudi v drugih gospodarskih panogah, kot je na primer trženje, kjer je uspešnost komunikacije odvisna tudi od slovnične ustreznosti besedil.

2. Namen in zgradba prispevka

Ob splošni razširjenosti pisarniških paketov za urejanje besedil do vse večjega izraza prihajajo tudi orodja za slovnično pregledovanje. Slednja uporabljajo tako poklicni kot tudi drugi uporabniki jezika. Med njimi so tudi šolarji in študentje, ki ta orodja pogosto uporabljajo kot obliž za pomanjkljivo slovnično in pravopisno znanje (Wei in Davies, 1997). Tovrstni načini uporabe teh orodij odpirajo tehtno vprašanje, v kolikšni meri je mogoče slovničnim pregledovalnikom zaupati in kdaj je njihova uporaba smiselna oziroma kdaj postane moteča. Zato smo se v tem prispevku osredotočili predvsem na opazovanje programskih orodij za slovnično pregledovanje v slovenščini, Besane in LanguageToola, z vidika njune natančnosti, zanesljivosti in praktične uporabnosti.

V sorodnih raziskavah s področja proučevanja slovničnih pregledovalnikov je mogoče opaziti, da so pogosto v ospredju nematerni govorci jezika. Iz tega razloga se nam je zdelo smiselno preveriti tudi učinkovitost tovrstnih orodij z vidika različnih ravni jezikovega znanja uporabnika. V ta namen smo oblikovali hipotezo, da uporabniki obravnavanih slovničnih pregledovalnikov pogosteje zaupajo predlaganim popravkom pri prevajanju v tuji jezik kot pri prevajanju v materni jezik. Z veliko verjetnostjo lahko namreč trdimo, da so uporabniki jezika pri izražanju v tujem jeziku manj suvereni kot v maternem jeziku. Prav tako so domnevno dojemljivejši za upoštevanje popravkov, ki jim jih predlaga navidezno zanesljiva avtoriteta, kot je orodje za slovnično pregledovanje.

V prispevku je predstavljen del obsežne magistrske raziskave o uporabi slovničnih pregledovalnikov za

slovenščino, Besane in LanguageToola, z vidika natančnosti, zanesljivosti in praktične uporabnosti. Prav tako so predstavljeni tudi izsledki preverjanja hipoteze, da uporabniki slovničnih pregledovalnikov pogosteje zaupajo predlaganim popravkom teh orodij pri prevajanju v tuji jezik kot pri prevajanju v materni jezik. Zaradi prostorske omejitve so bili v prispevek vključeni samo vsebinsko najpomembnejši sklopi te raziskave.[1]

V 3. razdelku sta opisana zasnova praktičnega poskusa in uporabljeno gradivo, s pomočjo katerih smo poskusili uresničiti postavljene raziskovalne cilje. V 4. razdelku smo obrazložili enega od možnih načinov za kvantitativno obdelavo izsledkov, pridobljenih pri tovrstnih raziskavah, in sicer z vidika parametrov natančnosti in zanesljivosti ter z vidika odnosa uporabnikov do predlaganih popravkov. Tej razlagi v 5. razdelku sledi preverjanje uvodoma postavljene hipoteze in predstavitev analize izsledkov, medtem ko smo v zaključku na podlagi zbranih ugotovitev podali predloge za izboljšave obravnavanih orodij in priporočila za nadaljnje raziskovalno delo na tem področju.

3. Praktični poskus

Pri opazovanju uporabe slovničnih pregledovalnikov za slovenščino z vidika omenjenih parametrov smo izhajali iz sorodnih raziskav o učinkovitosti tovrstnih orodij (Burston, 1996; Wei in Davies, 1997; Domeij, Knutsson in Severinson Eklundh, 2002). Na podlagi teh raziskav smo najprej opredelili zgoraj navedene raziskovalne cilje. Nato smo v skladu s cilji zasnovali praktični poskus, v okviru katerega smo udeležence poskusa prosili, naj v 90 minutah prevedejo dve 150 besed dolgi in tematsko manj zahtevni publicistični besedili. Prvo besedilo so udeleženci poskusa morali prevesti iz slovenščine v angleščino, drugo pa iz angleščine v slovenščino.

Skupino udeležencev poskusa je sestavljalo 21 študentov 1. letnika dodiplomskega študijskega programa Medjezikovno posredovanje na Oddelku za prevajalstvo Filozofske fakultete v Ljubljani, z jezikovno kombinacijo slovenščina-angleščina-nemščina. Izbira dane skupine je temeljila na predpostavkah, da imajo ti študentje v primerjavi s študenti višjih letnikov iste študijske smeri manj pravopisnega in slovničnega znanja ter manj izkušenj s prevajanjem besedil iz maternega v tuji jezik in obratno. S tem smo namreč želeli zagotoviti čim večji vzorec napak, na katerem smo pozneje preverjali natančnost in zanesljivost Besane in LanguageToola.

Izbira ustreznega gradiva za prevajanje je bila ključnega pomena za uspešno izvedbo poskusa, saj smo pričakovali, da bo število slovničnih napak, ki jih bodo udeleženci zagrešili, v veliki meri odvisno tako od pravopisne in slovnične kot tudi od pomenske in slogovne zahtevnosti izvirnega besedila. Prvo, slovensko izvirno besedilo, je bilo del daljše zgodbe, objavljene v reviji slovenskega letalskega prevoznika *Adria Airways*, medtem ko je bilo drugo, angleško izvirno besedilo, vzeto iz članka, objavljenega v spletni izdaji britanskega časnika *The Guardian*. Besedili sta se s tematskega vidika razlikovali do te mere, da se njuna nabora besedišča nista prekrivala (prim. Jurišić, 2013, Priloge: Izvirno besedilo_SLO/AN).

Ob tem je treba poudariti, da smo besedili za potrebe poskusa prilagodili tako, da smo določene dele besedila izbrisali, spremenili ali dodali, s čimer smo v oba izvirnika vnesli določene težje pravopisne in slovnične zagate (prim. Jurišić, 2013, Prilagojeni izvirni besedili_SLO/AN +Navodila za poskus). Te slovnične pasti smo črpali deloma iz nekaterih sorodnih raziskav o slovničnih pregledovalnikih, deloma pa iz gradiva za določene dodiplomske in podiplomske predmete pri študiju prevajanja na Oddelku za prevajalstvo (prim. Kies, 2008; Connors in Lunsford, 1992: 398; *Uvod v študij slovenskega jezika*, *Prevajalski seminar I* in *II – prevajanje iz slovenščine v angleščino/prevajanje iz angleščine v slovenščino*). Tudi s tem smo želeli povečati število morebitnih napak in tako zagotoviti čim reprezentativnejši vzorec za nadaljnjo analizo učinkovitosti obeh orodij.

Udeleženci poskusa so bili med poskusom razdeljeni v dve skupini, pri čemer je ena skupina prevajala v urejevalniku besedil *Microsoft Word 2010*, druga pa v urejevalniku besedil *Apache OpenOffice Writer 3.4.1*. Na ta način smo zagotovili, da je imela ena skupina udeležencev ob prevajanju v slovenščino na razpolago slovnični pregledovalnik *Amebis Besana*[2] *3.34*, druga pa sorodno odprtokodno orodje *LanguageTool 2.0*.[3] Obe orodji sta v tem primeru delovali v obliki programskih dodatkov k urejevalnikoma besedil, kar se je izkazalo za pomanjkljivost, ki jo bomo izpostavili v 5. razdelku.

Ključni element opisanega praktičnega poskusa je predstavljalo snemanje zaslona vseh udeležencev poskusa – brez njihove vednosti – s pomočjo snemalnika zaslona *TechSmith SnagIt 11.0.0*.[4] S tem smo pridobili vpogled tako v dejansko uporabo posameznega slovničnega pregledovalnika kot tudi v sam prevajalski proces vsakega udeleženca poskusa. Med obdelavo posnetkov smo namreč lahko opazovali, katere vrste napak je orodje zaznalo in katere spregledalo, katere predloge popravkov je zanje podalo ter kako so se udeleženci poskusa odzivali na omenjene zaznave napak in predlagane popravke pri posameznem slovničnem pregledovalniku. Pridobljene izsledke smo s pomočjo tipologije posameznih kategorij napak, podrobneje razložene v naslednjem razdelku, in opažanj, zabeleženih med pregledovanjem posnetkov, obdelali kvantitativno in kvalitativno.

[1] Celotno besedilo je na voljo na spletnem naslovu http://jurisicm.webs.com/magistrska_jurisic.pdf.

[2] *Amebis Besana* je programski paket, ki ga je razvilo podjetje Amebis in je namenjen odkrivanju pravopisnih in slovničnih napak v slovenskih besedilih (Holozan, 2012: 101; http://besana.amebis.si/).

[3] *LanguageTool* je odprtokodni slogovni in slovnični pregledovalnik, ki ga je leta 2003 v okviru svoje magistrske raziskave razvil Daniel Naber in trenutno podpira pregledovanje besedil v 29 jezikih (Naber, 2003: 3; https://languagetool.org/).

[4] http://www.techsmith.com/snagit.html

4. Tipologija posameznih kategorij napak za obdelavo izsledkov poskusa

Zasnova te tipologije napak ob upoštevanju ciljev raziskave temelji na opazovanju uporabe slovničnih pregledovalnikov z vidika natančnosti, zanesljivosti in odnosa uporabnikov do predlaganih popravkov. V tipologiji uporabljeno razlikovanje med mehanskimi in slovničnimi napakami smo uvedli na podlagi sorodnih raziskav o uporabi orodij za slovnično pregledovanje.

Med mehanske napake smo uvrstili vse napake v črkovanju, saj odkrivanje tovrstnih napak poteka na podlagi preverjanja posameznih besed s pomočjo enostavnih algoritmov iskanja in ujemanja (Trost, 2004: 37; Voutilainen, 2004: 228). Nasprotno pa preverjanje slovnične ustreznosti temelji na preverjanju skupine besed, ki so med seboj povezane na različne načine in med katerimi veljajo zapletene slovnične zakonitosti, zato razvoj tovrstnih orodij od razvijalca zahteva dobro poznavanje slovnice in metod za njeno formalizacijo (Burston, 1996: 106). Zato je mogoče trditi, da preverjanje črkovanja v besedilu poteka na mehanski ravni, medtem ko slovnično preverjanje poteka na višji, zahtevnejši ravni obdelave naravnega jezika.

Kot temeljne kategorije opazovanja smo opredelili *Zaznavo napak*, *Zanesljivost nasvetov in popravkov* ter *Odnos do nasvetov in popravkov* (prim. Tabela 1 in 2). Ob tem je treba opozoriti, da smo najprej razvili zgolj ogrodje tipologije napak – temeljne kategorije in podkategorije –, medtem ko smo posamezne vrste napak dodajali postopoma in po potrebi med samo analizo izsledkov (prim. Jurišić, 2013, Priloge: Vrste posameznih mehanskih in slovničnih napak_SLO; Vrste posameznih mehanskih in slovničnih napak_AN). Poudariti je treba tudi to, da vse opredelitve temeljnih kategorij in podkategorij z izjemo posameznih vrst napak veljajo za celotno obdelavo izsledkov, ki se nanaša tako na angleške kot tudi slovenske primere napak.

	Mehanske napake	**Slovnične napake**
Zaznava napak	Ustrezno zaznane Napačno zaznane Nezaznane	
Zanesljivost nasvetov in popravkov	Pravilen popravek Napačen vir napake Pravilen popravek (sprejet) Napačen popravek (sprejet)	Uporaben nasvet Nejasen nasvet Napačen vir napake Uporaben nasvet (sprejet) Napačen nasvet (sprejet)
Odnos do nasvetov in popravkov	Končni pregled napak Posredno upoštevanje Neposredno upoštevanje	

Tabela 1: Delitev temeljnih kategorij in podkategorij obdelave izsledkov

Podkategorije določenih temeljnih kategorij napak se v veliki meri prekrivajo, s čimer smo želeli zagotoviti čim večjo enotnost in doslednost pri analizi izsledkov. Na obeh ravneh smo namreč proučevali napake, ključna razlika v njihovi obravnavi pa je razvidna iz načina zaznave napak na posamezni ravni. Obe orodji, Besana in LanguageTool, mehansko napako (npr. napačen zapis besede) označita takoj, ko uporabnik konča s pisanjem dane besede in pritisne preslednico, medtem ko slovnično napako (npr. manjkajočo vejico) praviloma podčrtata šele potem, ko uporabnik poved dokonča s končnim ločilom in nadaljuje s pisanjem nove enote besedila.

Skladno s tovrstnimi zakonitostmi delovanja posameznega slovničnega pregledovalnika smo oblikovali tudi pristop k razvrščanju primerov napak v posamezne kategorije in podkategorije. Ustrezno zaznane mehanske napake smo denimo lahko zabeležili takoj po uporabnikovem pritisku na preslednico, medtem ko smo pri zapisovanju slovničnih napak morali vedno počakati, da je uporabnik poved dokončal, saj ga je program šele tedaj opozoril na obstoječo slovnično napako. Če je uporabnik napako pred opozorilom že popravil, potem te napake nismo zabeležili. Podobno v sklopu podkategorij *Nezaznane mehanske/slovnične napake* nismo beležili napak, ki jih tovrstna orodja še ne morejo zaznati, temveč zgolj tiste, ki so dejansko – ali vsaj teoretično – znotraj njihovega dosega. Prav tako pri zaznavi napak in odnosu do popravkov nismo zapisovali popravkov, ki jih je urejevalnik Word izvedel v okviru funkcije *Samopopravki*, saj večina uporabnikov teh popravkov ni niti opazila.

Uporabo dane tipologije je lažje ponazoriti s pomočjo primera, zato smo v nadaljevanju navedli dva primera razvrščanja napak. V prvem, enostavnem primeru se je udeleženec med prevajanjem v slovenščino zatipkal in namesto besede »pričakovati« zapisal »špričakovati«. Program je tipkarsko napako nemudoma prepoznal in podčrtal, udeleženec pa je z desnim klikom besede odprl ustrezen predlog popravka ter z levim klikom zatipkano besedo zamenjal z ustreznim zapisom. Ta izsek njegovega prevajalskega procesa smo na podlagi opisane tipologije obdelali v 4 logičnih korakih:

1. Udeleženec poskusa je besedo napačno zapisal, torej govorimo o mehanski napaki (vrsta napake: »napačen zapis besede (zatipkano)«).
2. Program je to mehansko napako zaznal, kar je razvidno iz podčrtave, ustreznost zaznave pa potrjuje predlog popravka.
3. Program je za to ustrezno zaznano mehansko napako podal pravilen predlog popravka (druga

možnost bi bila, da orodje zaradi napačnega vira napake izpiše neustrezen predlog popravka).

4. Z levim klikom ponujenega predloga je udeleženec popravek sprejel, torej lahko v zadnjem koraku zabeležimo, da je udeleženec neposredno upošteval predlog popravka za dano mehansko napako.

Drugi primer obdelave je težavnejši, saj smo se morali odločiti med uvrščanjem napake v eno od navidezno podobnih podkategorij *Napačne zaznave napake* in *Napačnega vira napake*. Udeleženec poskusa je med prevajanjem izpustil piko pri navedbi datuma v slovenščini (primer: »Ljubljana, 17[.] **september** 2004«; krepki tisk označuje mesto napake, oglati oklepaji pa vsebujejo manjkajoče znake). Program je to mehansko napako zaznal, vendar jo je opredelil kot »neujemanje s pridevnikom«. To pomeni, da je program sicer ustrezno prepoznal obstoj določene vrste napake, vendar je napačno določil njen izvor. Na podlagi tega smo dano napako opredelili kot *Ustrezno zaznavo* in *Napačen vir napake*.

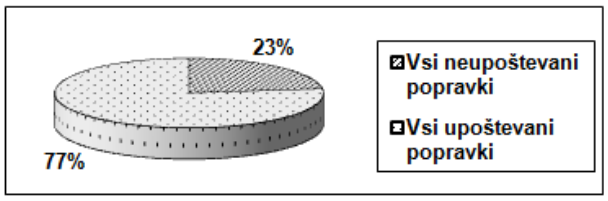
Zaznava napak	
Ustrezno zaznane mehanske/slovnične napake	Orodje v dani enoti besedila prepozna ustrezno vrsto mehanske/slovnične napake.
Napačno zaznane mehanske/slovnične napake	Orodje v dani enoti besedila, ki ne vsebuje napak, prepozna določeno vrsto mehanske/slovnične napake (lažni alarm).
Nezaznane mehanske/slovnične napake	Orodje v dani enoti besedila ne prepozna obstoječe mehanske/slovnične napake.
Nasveti in popravki (mehanske napake)	
Pravilen popravek	Orodje predlaga pravilen popravek za ustrezno zaznano mehansko napako.
Napačen popravek	Orodje zaradi napačne zaznave mehanske napake (lažni alarm) predlaga napačen popravek.
Napačen VIR napake	Orodje zazna obstoj določene mehanske napake, vendar poda predlog popravka za napačno vrsto napake.
Pravilen popravek (sprejet)	Sprejme pravilen popravek za ustrezno zaznano mehansko napako (popravi sam ali s pomočjo orodja).
Napačen popravek (sprejet)	Sprejme napačen popravek zaradi napačnega vira napake ali napačne zaznave (popravi sam ali s pomočjo orodja).
Nasveti in popravki (slovnične napake)	
Uporaben nasvet	Orodje poda uporaben nasvet, ki omogoča pravilno popravljanje ustrezno zaznane slovnične napake.
Nejasen nasvet	Orodje poda nejasen nasvet, ki oteži pravilno popravljanje ustrezno zaznane slovnične napake.
Napačen nasvet	Orodje zaradi napačne zaznave slovnične napake (lažni alarm) poda napačen nasvet.
Napačen VIR napake	Orodje zazna obstoj določene slovnične napake, vendar poda nasvet za napačno vrsto napake.
Uporaben nasvet (sprejet)	Sprejme uporaben nasvet za ustrezno zaznano slovnično napako (popravi sam ali s pomočjo orodja).
Napačen nasvet (sprejet)	Sprejme napačen nasvet zaradi napačnega vira napake ali napačne zaznave (popravi sam ali s pomočjo orodja).
Odnos do nasvetov in popravkov	
Neupoštevanje popravkov (neznana beseda)	Ne upošteva predloga popravka za neznano besedo (lastno ime, kraj ipd.).
Ni končnega pregleda napak (število primerov)	Ne pregleda zaznanih napak ob koncu prevajanja (v kolikšnem številu primerov?).
Posredno upoštevanje mehanskih/slovničnih popravkov	Posredno upošteva predlog popravka za ustrezno zaznano mehansko/slovnično napako, pri čemer sam vnese popravek.
Neposredno upoštevanje mehanskih/slovničnih popravkov	Neposredno upošteva predlog popravka za ustrezno zaznano mehansko/slovnično napako, pri čemer orodje vnese popravek.

Tabela 2: Razlage podkategorij po sklopih posameznih kategorij

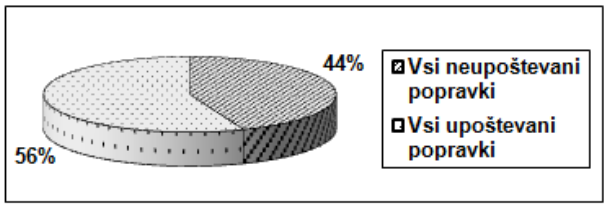
5. Analiza izsledkov

Na podlagi izsledkov, pridobljenih med izvedbo praktičnega poskusa, smo najprej preverili veljavnost hipoteze, da uporabniki slovničnih pregledovalnikov pogosteje upoštevajo predlagane popravke teh orodij pri prevajanju v tuji jezik v primerjavi s prevajanjem v materni jezik. Ob tem smo opazovali razmerje med številom popravkov, ki so jih udeleženci poskusa v času

prevajanja pri posameznem jeziku upoštevali, in številom vseh popravkov, ki so jih v poskusu uporabljena orodja ponudila. To razmerje smo ponazorili z naslednjima grafikonoma (N v vseh prikazih označuje velikost vzorca):

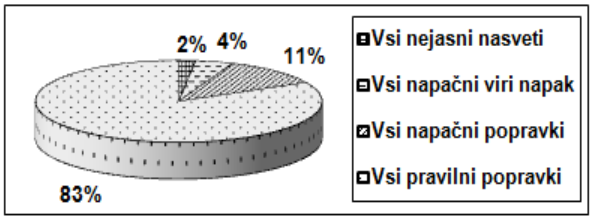


Grafikon 1: Upoštevanje popravkov pri prevajanju v angleščino v odstotkih (N=212)

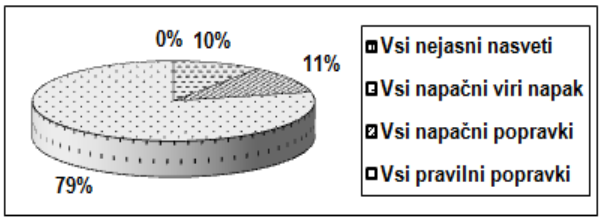


Grafikon 2: Upoštevanje popravkov pri prevajanju v slovenščino v odstotkih (N=177)

Statistični podatki iz pričujočih grafikonov so potrdili veljavnost omenjene hipoteze, saj so udeleženci poskusa pri prevajanju v angleščino v dobrih 20 % več primerov upoštevali predloge popravkov kot pri prevajanju v slovenščino. Tudi deleža napačnih popravkov sta bila pri obeh jezikih primerljiva, s čimer smo ovrgli morebitne dvome, da je stopnja sprejemanja popravkov pri določenem jeziku nižja zaradi večjega števila neustreznih predlogov popravkov (prim. Grafikon 3 in 4).



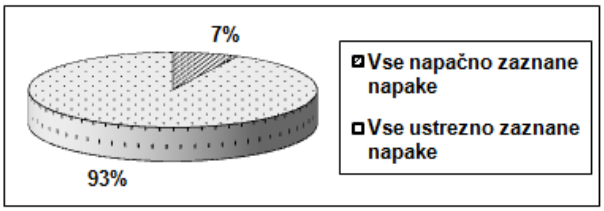
Grafikon 3: Delitev predlaganih popravkov glede na njihovo ustreznost pri prevajanju v angleščino (N=212)



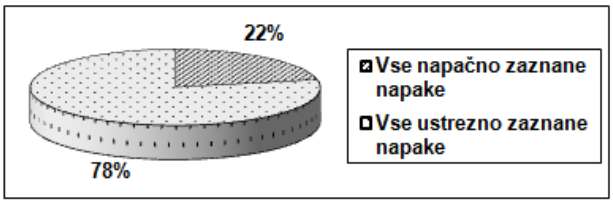
Grafikon 4: Delitev predlaganih popravkov glede na njihovo ustreznost pri prevajanju v slovenščino (N=177)

Pri proučevanju uporabe slovničnih pregledovalnikov se kot objektivni merili za ugotavljanje učinkovitosti tovrstnih orodij najpogosteje uporabljata natančnost in priklic. Parameter natančnosti lahko opredelimo kot razmerje med ustrezno zaznanimi napakami in vsemi v besedilu zaznanimi napakami, medtem ko parameter priklica podaja razmerje med zaznanimi napakami in vsemi obstoječimi napakami v besedilu (Domeij, Knutsson in Severinson Eklundh, 2002: 263). Izmerjene vrednosti obeh parametrov so prikazane v grafikonih 5 in 6 (prim. Grafikon 5 in 6).

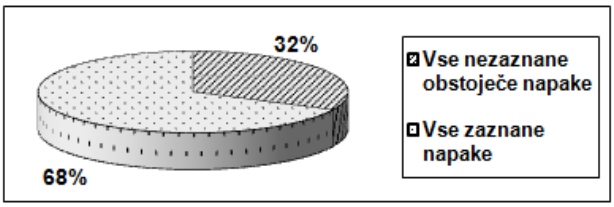
Iz teh grafikonov je razvidno, da je bila Besana pri ustreznem popravljanju napak v danem primeru bistveno natančnejša od LanguageToola, saj je ta v primerjavi z Besano zaznane napake ustrezno popravil v 15 % manj primerov. Na ravni odkrivanja napak med obema orodjema ni bilo opaziti večjih razlik, saj sta obe zaznali dobri dve tretjini vseh napak, ki so jih vsebovali prevodi udeležencev (prim. Grafikon 7 in 8).



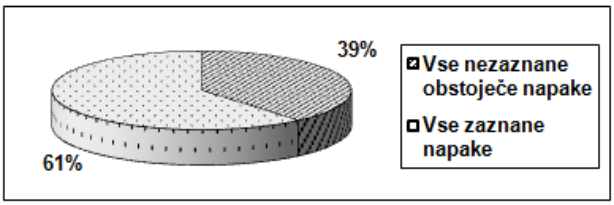
Grafikon 5: Natančnost slovničnega pregledovalnika Besana v odstotkih (N=122)



Grafikon 6: Natančnost slovničnega pregledovalnika LanguageTool v odstotkih (N=55)



Grafikon 7: Priklic slovničnega pregledovalnika Besana v odstotkih (N=180)



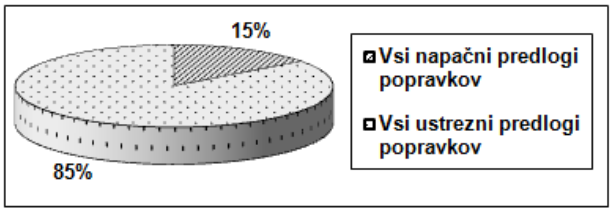
Grafikon 8: Priklic slovničnega pregledovalnika LanguageTool v odstotkih (N=90)

Razliko v natančnosti in priklicu med obema pregledovalnikoma je mogoče pripisati temu, da je Besana kot plačljivo orodje pri razvoju deležna precej večje finančne in strokovne podpore kot prosto dostopno odprtokodno orodje LanguageTool. Poleg tega je mogoče domnevati, da bi LanguageTool ob večjem vzorcu napak verjetno pridobil nekaj več odstotkov na ravni natančnosti in priklica.

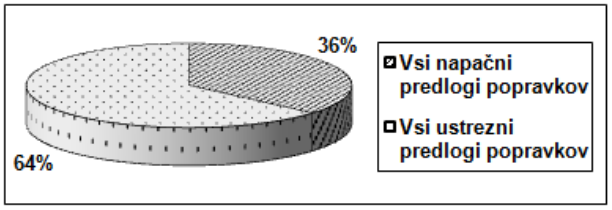
Čeprav je slovnične pregledovalnike zaradi razlik v slovničnih zakonitostih med jeziki, sestavi in velikosti vzorca ter stopnji zahtevnosti posameznih napak težko primerjati med seboj, je vseeno mogoče trditi, da tako

Besana kot tudi LanguageTool v primerjavi s tovrstnimi orodji za druge jezike dosegata zadovoljivo raven natančnosti in priklica. Slovnični pregledovalnik BonPatron za francoščino ima na primer 91 % natančnost in 88 % priklic, Granska za švedščino 82 % natančnost in 63 % priklic, medtem ko ima slovnični pregledovalnik za švedščino v Wordu 2000 le 47 % natančnost in 66 % priklic (Burston, 2008: 341; Carlberger in dr., 2004: 14).

Parameter zanesljivosti smo pri opazovanju uporabe omenjenih slovničnih pregledovalnikov za slovenščino opredelili kot razmerje med predlogi popravkov, ki uporabnika tako posredno kot tudi neposredno vodijo do ustrezne rešitve, in vsemi predlogi popravkov, ki jih je posamezno orodje ponudilo (prim. Grafikon 9 in 10).



Grafikon 9: Zanesljivost slovničnega pregledovalnika Besana v odstotkih (N=122)



Grafikon 10: Zanesljivost slovničnega pregledovalnika LanguageTool v odstotkih (N=55)

Izsledki te primerjave so pokazali, da je bila Besana v tem primeru precej zanesljivejša od LanguageToola, saj je napake uspešno zaznala in popravila v dobrih 20 % primerov več kot LanguageTool. Možne razloge smo navedli v predhodnem odstavku.

Praktično uporabnost obeh slovničnih pregledovalnikov za slovenščino, Besane in LanguageToola, smo proučevali med obdelavo izsledkov poskusa in preverjanjem načinov delovanja različnih programskih oblik obeh orodij, dodatne informacije pa smo pridobili tudi s pomočjo intervjujev z glavnima razvijalcema obeh orodij.

Obe orodji je mogoče uporabljati v obliki dodatkov k urejevalnikom besedil ali kot samostojna programa, pri čemer imata oba načina uporabe svoje prednosti in slabosti. Pri obeh orodjih je namreč raven natančnosti in zanesljivosti višja, če ju uporabljamo kot samostojna programa, saj se stavčna analiza, na kateri temelji preverjanje, v tem primeru izvede na dokončanih povedih (Jurišić, 2013: 37). Med pisanjem besedil v urejevalniku orodji slovnično ustreznost preverjata sproti – na nedokončanih povedih –, zato je število možnih interpretacij posameznih kombinacij besed neprimerno večje, kar zmanjša natančnost analize in s tem tudi ustrezno zaznavo napak (Jurišić, 2013: prav tam). Kljub temu je sprotno preverjanje slovnice za uporabnika časovno ugodnejše, saj je pri uporabi orodja v obliki samostojnega programa popravke v besedilo treba vnesti ročno, kar je pri daljšem besedilu precej zamudno.

Ena od vidnejših šibkosti obeh orodij je tudi v tem, da lahko slovnično manj veščega uporabnika hitro zavedejo z lažnim alarmom (to pomeni, da orodje javi napako v popolnoma pravilni povedi). Ti primeri so pogosto vodili v dejanske napake, ki so jih uporabniki zagrešili med tem, ko so poskušali popraviti domnevno napako. LanguageTool je na primer v stavku »saj je **kar** 27 % prebivalcev univerzitetnih študentov« napačno predlagal manjkajočo vejico, ker je členek »kar« interpretiral kot veznik. Udeleženec poskusa je njegov predlog popravka upošteval in s tem storil dejansko napako. Kljub omenjenim šibkostim na ravni delovanja in uporabe posameznih programskih oblik obeh orodij je mogoče trditi, da uporaba obeh orodij udeležencem poskusa ni povzročala večjih težav, temveč je v številnih primerih pripomogla k višji kakovosti njihovega prevajalskega dela.

6. Zaključek

V prispevku smo predstavili analizo uporabe slovničnih pregledovalnikov za slovenščino, Besane in LanguageToola, z vidika parametrov natančnosti, zanesljivosti in praktične uporabnosti. Ob tem se je Besana izkazala za natančnejšo in zanesljivejšo od LanguageToola, medtem ko na ravni praktične uporabnosti med obema orodjema ni bilo mogoče zaznati bistvenih razlik.

Preverili in potrdili smo hipotezo, da uporabniki obeh obravnavanih orodij pogosteje zaupajo predlaganim popravkom pri prevajanju v tuji jezik v primerjavi s prevajanjem v materni jezik. Poleg tega smo razvili tipologijo posameznih kategorij napak, ki je lahko v pomoč pri nadaljnjih raziskavah o uporabi slovničnih pregledovalnikov z vidika empirično lažje ali težje izmerljivih parametrov.

Obravnavana slovnična pregledovalnika bi bilo mogoče izboljšati s prilagajanjem njunih pristopov k prepoznavanju napak in predlaganju popravkov za uporabnike z različnimi ravni jezikovnega znanja (na primer manjša natančnost in večji priklic za materne govorce jezika in obratno za nematerne govorce). Prav tako bi metodo izvedbe poskusa lahko nadgradili z uvedbo vprašalnikov za udeležence, s čimer bi parametre opazovanja, kot sta na primer odnos uporabnikov do popravkov in praktična uporabnost, dopolnili z dejanskimi uporabniškimi izkušnjami. Ob koncu želimo izpostaviti, da je metodo snemanja zaslona mogoče uporabiti tudi za proučevanje dejanskega prevajalskega procesa. Ob tem je na primer mogoče opazovati iskanje prevodnih ustreznic pri težjih delih izvirnika in načine uporabe slovarjev ter s tem povezano raven informacijske pismenosti v povezavi s hitrostjo in kakovostjo prevajanja.

7. Literatura

Amebis Besana - Datoteka s pomočjo. Priloga programskemu paketu Amebis Besana. (Dostop 26. 3. 2013)

Burston, J., 1996. A comparative evaluation of French grammar checkers. *CALICO Journal*. 13/2–3. 104–111.

Burston, J., 2008. BonPatron: An Online Spelling, Grammar, and Expression Checker. *CALICO Journal*. 25/2. 337–347.

Carlberger, J., Domeij, R., Kann, V. in Knutsson, O., 2004. The development and performance of a grammar checker for Swedish: A language engineering perspective. *Natural Language Engineering*. 1/1. 1–17.

Connors, R. in Lunsford, A., 1992. Frequency of Formal Errors in Current College Writing, or Ma and Pa Kettle Do Research. Connors, R. in Glenn, C., (ur.): *The St. Martin's Guide to Teaching Writing*. 2. izdaja. New York: St. Martin's. Izvorno objavljeno v: Connors, R. in Lunsford, A., 1988: Frequency of Formal Errors in Current College Writing, or Ma and Pa Kettle Do Research. *College Composition and Communication*. 39/4. 395–409.

Domeij, R., Knutsson, O. in Severinson Eklundh, K., 2002. Different ways of evaluating a Swedish grammar checker. *Proceedings of the 3rd International Conference Language Resources and Evaluation (LREC 2002)*. Las Palmas, Španija. 262–267.

Helfrich, A. in Music, B., 2000. Design and evaluation of grammar checkers in multiple languages. *COLING '00 Proceedings of the 18th conference on Computational linguistics*. Zvezek 2. Stroudsburg, Pensilvanija, ZDA: Association for Computational Linguistics. 1036–1040.

Holozan, P., 2012. Kako dobro programi popravljajo vejice v slovenščini. Erjavec, T. in Žganec Gros, J., (ur.): *Zbornik Osme konference Jezikovne tehnologije, 8. do 12. oktober 2012, [Ljubljana, Slovenia] : zbornik 15. mednarodne multikonference Informacijska družba - IS 2012. zvezek C*. Ljubljana: Inštitut Jožef Štefan. 101–106.

Holozan, P., 2013. Osebni pogovor z dne 24. 1. 2013. Kamnik.

Jurišić, M., 2013. *Slovnični pregledovalniki za slovenščino – pregled in uporaba*. Magistrsko delo, Filozofska fakulteta, Univerza v Ljubljani.

Kies, D., 2008. Evaluating Grammar Checkers. A Comparative Ten-Year Study. *Proceedings of the 6th International Conference on Education and Information Systems, Technologies and Applications: EISTA 2008*. Orlando, Florida, ZDA. http://papyr.com/hypertextbooks/grammar/gramchek.htm. (Dostop 26. 2. 2013)

Naber, D., 2003. *A Rule-based Style and Grammar Checker*. Bielefeld: Technische Fakultät, Universität Bielefeld.

Trost, H., 2004. Morphology. Mitkov, R., (ur.), *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press. 25–47.

Voutilainen, A., 2004. Part-of-speech Tagging. Mitkov, R., (ur.), *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press. 219–239.

Wei, Y. H. in Davies, G., 1997: Do grammar checkers work? Kohn, J., Rüschoff, B. in Wolff, D., (ur.), *New horizons in CALL: Proceedings of EUROCALL 96*. Szombathely: Daniel Berzsenyi College. http://www.camsoftpartners.co.uk/euro96b.htm. (Dostop 21. 12. 2012)

SecondEGO – virtualni pomočnik za vsakogar

Miro Romih

Amebis, d. o. o., Kamnik
Bakovnik 3, 1241 Kamnik
miro.romih@amebis.si

Povzetek

Prispevek predstavlja sistem SecondEGO, ki je namenjen izdelavi lastnih virtualnih pomočnikov. Ciljna publika sistema so vsi tisti, ki želijo na svojih spletnih straneh uporabnikom ponuditi pomoč v obliki direktnih odgovorov na njihova vprašanja. Ti odgovori se generirajo tudi ob pomoči jezikovnih tehnologij, možno pa je vključevati tudi naprednejše funkcije.

SecondEGO – virtual agent for everyone

This paper presents the SecondEGO system, which is designed to create your own virtual agents. The target audience of the system are all those who wish to offer assistance in the form of direct answers to user's questions on their websites. These answers are generated with the help of language technologies, but it is also possible to include other advanced functions.

1. Uvod

Običajno imamo na voljo več načinov, kako priti do informacije. Vzemimo na primer, da smo na avtobusni postaji in želimo izvedeti, kdaj odpelje prvi avtobus v nek kraj. Na zidu je obešen vozni red, poleg njega pa je okence za informacije. Nekateri si bodo čas odhoda prebrali iz voznega reda, drugi pa bodo odgovor iskali pri okencu za informacije.

Podobno je na spletu. Določeno informacijo na neki spletni strani lahko poiščemo sami, lahko pa, če obstaja seveda, preko pogovornega okenca o tem povprašamo operaterja, ki na naše vprašanje čaka v ozadju. Če operater (izven delovnega časa) ni dostopen, lahko njegovo vlogo prevzame virtualni pomočnik, ki je sposoben odgovarjati vsaj na najpogostejša vprašanja uporabnikov.

2. Namen članka

Namena članka je predstavitev sistema SecondEGO, ki ga je razvilo podjetje Amebis. Dostopen je na naslovu www.secondego.com, omogoča pa izdelavo lastnih virtualnih pomočnikov na enostaven in hiter način.

Poleg samega (vidnega dela) portala je predstavljeno tudi njegovo ozadje, ki vključuje nekatere zanimive jezikovne tehnologije.

3. Virtualni pomočniki

Virtualni pomočniki, kakor jih poimenujem v tem prispevku, so programi, zasnovani za konverzacijo z uporabniki. Imajo številna imena – virtualni asistenti, navidezni agenti, programirani sogovorniki itd. V angleščini obstaja nekaj sto izrazov, ki opisujejo te programe oz. sisteme za različne namene (npr. »virtual agent«, »virtual assistant«, za klepetanje »chatbot« ali »bot«).

Tudi sistemov oz. platform, ki podpirajo to funkcionalnost, je precej, z njihovo pomočjo pa danes deluje na tisoče bolj ali manj uporabnih virtualnih pomočnikov. Med bolj znanimi starejšimi in še danes delujočimi sta Eliza (Weizenbaum, 1966) in Alice (Wallace, 2003), med sodobnimi pa izstopata Siri (Apple) in Watson (IBM). Vsak od omenjenih predstavlja tudi nov mejnik tehnologije za komunikacijo oz. dialog v naravnem jeziku.

Vsak virtualni pomočnik je običajno namenjen določeni ciljni skupini in odgovarja na vprašanja z bolj ali manj omejenega področja. Največkrat se jih uporablja za podporo uporabnikom (tudi kot neke vrste »pametni« iskalnik), kot marketinško orodje (za večji odziv uporabnikov), za zabavo (klepetanje), ali kot dodatek uporabniškemu vmesniku za lažjo, hitrejšo in naravnejšo komunikacijo.

Kot je bilo že omenjeno, obstaja veliko število sistemov, s katerimi je mogoče graditi virtualne pomočnike. Večina je zelo enostavnih, in omogoča dialog v poljubnem naravnem jeziku. Je pa za slovenski jezik treba v takem sistemu ustvariti bazo vprašanj in odgovorov povsem na novo, saj za zdaj še ni na voljo nobene baze, ki bi jo lahko uporabili kot osnovo. Izgradnja take baze namreč zahteva precej truda, če želimo, da bo osnovno »znanje« kolikor toliko zadovoljivo. Še posebej za pregibni jezik, kot je naš, saj je v teh sistemih vse potrebne besedne oblike za zdaj treba vpisovati ročno.

Naprednejši sistemi sicer imajo vgrajeno boljšo jezikovno podporo, ampak največkrat le za angleščino, sem in tja še za kak drug svetovni jezik, za slovenščino pa te podpore trenutno ni.

Za izdelavo virtualnega pomočnika, ki bi se znal pogovarjati v slovenskem jeziku, smo tako obsojeni na uporabo enostavnih mehanizmov vnosa znanja, ki zahtevajo veliko ročnega dela, ali pa razviti sistem, ki bi imel vgrajene mehanizme za obvladovanje slovenščine.

V Amebisu smo se odločili za drugo možnost in razvili svoj lastni sistem, v katerega lahko poljubno dodajamo različne module. Enostavne/osnovne, specifične za določenega uporabnika, pa tudi nekatere take, ki so bili razviti za druge sisteme in omogočajo uporabo že razvitih baz znanja za tuje jezike.

4. SecondEGO

Sistem SecondEGO je sestavljen iz jedra in vmesnika/portala, s katerim je sistem mogoče upravljati. Z njim lahko uporabniki enostavno in hitro kreirajo svoje virtualne pomočnike, jih učijo, testirajo in analizirajo njihovo delovanje ter uporabo.

Slika1. Grafična podoba sistema SecondEGO

4.1. Jedro

SecondEGO je vzorčno voden sistem. Deluje na osnovi vzorcev, v katerih so odgovori na vnaprej določena vprašanja ali ključne besede. Sistem postavljeno vprašanje primerja z vsemi vzorci in izbere tistega, ki se po ključnih besedah najbolj ujema. Tako v osnovi delujejo praktično vsi sistemi za uporabo virtualnih pomočnikov.

Prva značilnost, ki sistem SecondEGO razlikuje od večine primerljivih rešitev, je modularna zgradba. To pomeni, da so vzorci v posameznih modulih lahko zapisani v več različnih formatih, ki jih sistem podpira. Modulov določenega formata je lahko poljubno veliko, vsak modul pa lahko vsebuje poljubno število vzorcev.

Druga razlika je, da se za izbrane jezike vhodno vprašanje najprej morfološko, sintaktično in semantično analizira. Na osnovi rezultatov te analize je veliko več možnosti za izbiro ustreznega odgovora, seveda pa morajo vzorci omogočati zapis teh informacij. Zato smo razvili poseben skriptni jezik. Gre za podoben jezik, kot je ChatScript (ChatScript Open Source project), le da je njegova sintaksa drugačna, saj smo ga razvili že precej let prej, vanj pa je vgrajena podpora za slovenski in angleški jezik. Z večjim dodatnim vložkom je mogoče vgraditi podporo tudi za druge jezike.

V tem jeziku so napisani vsi vgrajeni moduli znanja in nekaj specifičnih modulov za določene uporabnike (npr. virtualna računovodkinja Zdenka podjetja Pronet Kranj).

Vnos znanja na ta način je veliko hitrejši, omogoča vnos lem, pomenov, uporabo referenčnih datotek, vključitev zunanjih podatkovnih baz, ter poljubnega števila lastno sprogramiranih novih funkcij. Z njihovo pomočjo lahko realiziramo praktično vse, kar uporabniki zahtevajo od virtualnega pomočnika.

Vzorci so lahko zelo enostavni in lovijo točno določen niz besed (slika2) ali določene ključne besede (slika3). Lahko pa namesto besed na osnovi pomenske analize vprašanja in v skriptni jezik vgrajene obsežne baze pomenov, povezanih v mrežo nadpomenov, podpomenov in drugih pomenskih relacij, v vzorcih uporabljamo tudi naprednejše funkcije (slika4).

```
$ (25) dobro odgovarjaš > Hvala.
```

Slika2. Enostavni vzorec

```
$ @KljucneBesede("nič|ničesar", "ne", "veš")
>
[_tip == 0] Pa me poskusite vprašati drugače.
[_tip == 1] To pa ni res.
```

Slika3. Lovljenje ključnih besed

```
$ @__JeBeseda("kaj je|kva je|kuga je|ka
je|kaj so|kva so|kuga so|ka so|kdo je|kdo
so|kaj veš o|kva veš o|kuga veš o|ka veš
o|poznaš|ali poznaš|a poznaš|al poznaš")
@OsnovniNadpomen([[zdravnik{0:0:0}]])
>
To je vrsta zdravnika. Podobno kot
@IzberiPodpomen([[zdravnik{0:0:0}]], #1)[So--
i---e].
```

Slika4. Uporaba nadpomenov in podpomenov

Tako se npr. z uporabo lem in pomenov v vzorcih lahko znebimo ročnega vnašanja besedni oblik, če to želimo, na drugi strani pa lahko uporabnik z uporabo poenostavljenih vzorcev enostavno in povsem zadovoljivo reši večino problemov, ki jih mora rešiti njegov virtualni pomočnik (slika5).

Slika5. Poenostavljena predstavitev vzorcev

Ena od zanimivih jezikovnotehnoloških funkcij, ki je vgrajena v sistem SecondEGO, je tudi obravnava oz. upoštevanje tipkarskih napak. Mnogokrat se namreč zgodi, da se ljudje pri pisanju vprašanja zatipkajo, potem pa se začudeno sprašujejo, kako to, da virtualni pomočnik ne zna odgovoriti na tako enostavno vprašanje. Svoje tipkarske napake velika večina seveda ne opazi, še kako pa ta zmoti funkcijo za primerjavo ključnih besed, ki zaradi zatipkane besede ne najde ustreznega vzorca. V sistemu SecondEGO smo primerjavo ključnih besed nadgradili tako, da se pri primerjavi vsake besede do neke mere upoštevajo tudi vse morebitne tipkarske napake. Zaradi časovne optimizacije seveda šele potem, ko v prvem prehodu ne najdemo točnih zadetkov. Taka funkcionalnost se je pokazala za zelo uporabno.

Ker je sistem zasnovan tako, da lahko vključujemo različne module, pripravljamo še dva, ki bosta pripomogla k dodatni uporabnosti sistema in pripomogla k dodatni konkurenčnosti v primerjavi z drugimi.

Prvi modul bo lahko vključeval zapis znanja v formatu AIML (Wallace, 2003, 2005; AIML), ki je standard na tem področju. Za različne jezike je na voljo kar nekaj že narejenih in celo prosto dostopnih baz znanja v tem formatu, ki jih bo mogoče enostavno vključiti. Za slovenščino po nam znanih podatkih s tem sicer ne bomo veliko pridobili, za večje svetovne jezike pa se bo baza osnovnega znanja na ta način hitro povečala.

Drugi modul, ki ga pripravljamo, bo omogočal direktno vključitev naše tehnologije Piflar na zelo enostaven način. Uporabnik bo vpisal le dejstva v obliki stavka ali odstavka, sistem pa bo na osnovi njihove analize sam odgovarjal na vprašanja v zvezi z njimi. Če bomo npr. vpisali stavek »Besana je avtomatska lektorica, ki odkriva slovnične napake v slovenskih besedilih.«, bo sistem znal odgovoriti na vprašanja »Kaj je Besana?«,

»Kaj dela/odkriva Besana?« ipd. Enako velja tudi za angleški jezik. Take tehnologije za zdaj nima še noben primerljiv sistem v svetu. Testno je ta tehnologija že vključena kot del iskanja po spletnih (pod)straneh, kjer pa v praksi še ne pride do polnega izraza.

4.2. Vmesnik

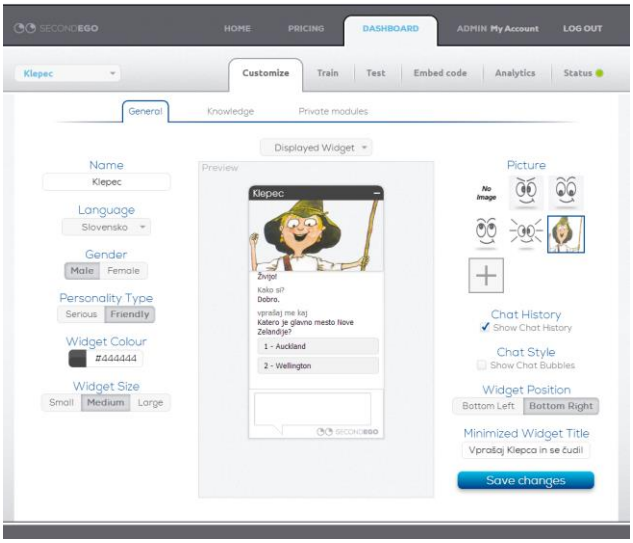
Ker problemi, ki jih sistem SecondEGO oz. z njim izdelani virtualni pomočniki rešujejo, niso omejeni le na slovenski prostor, je naš cilj tudi prodor na svetovni trg, zato je vmesnik portala v angleškem jeziku.

Uporaba portala je brezplačna, potrebno se je le registrirati. Plačljivost pride v poštev šele, ko dejansko število odgovorov nekega virtualnega pomočnika preseže določeno vrednost.

Po registraciji/prijavi ima uporabnik na voljo osnovne funkcionalnosti za upravljanje s pomočnikom.

4.2.1. Osnovne nastavitve

Ime virtualnega pomočnika je edini podatek, ki ga je potrebno vpisati. Vse ostale nastavitve si lahko uporabnik prilagodi glede na svoje cilje. Med te nastavitve sodijo jezik, spol, tip (uraden ali prijateljski), barva in velikost okenca, slika, prikaz zgodovine pogovora, oblika prikaza pogovora itd.



Slika6. Osnovne nastavitve virtualnega pomočnika

4.2.2. Osnovno znanje

Ko naredimo novega virtualnega pomočnika, ta seveda ne ve nič o specifičnih stvareh, na katere naj bi odgovarjal. Lahko pa mu za začetek dodamo nekaj splošnega znanja o stvareh, po katerih sogovorniki običajno povprašujejo. To znanje je trenutno zbrano v petih že vgrajenih modulih:

- »Osebnost« - modul vsebuje odgovore na najpogostejša vprašanja o osebnosti virtualnega pomočnika (»Kako ti je ime?«, »Ali rad bereš?«, »Ti si pameten.«, »Kaj najraje ješ?«, »Ali imaš kaj otrok?«);
- »Splošno znanje« - modul z vgrajenim splošnim znanjem (»Povej mi kakšno šalo.«, »Koliko je 3 + 2?«, »Sklanjaj samostalnik miza.«, »Kdaj bo pust?«, »Koliko je ura?«);
- »Wikipedija« - informacije iz Wikipedije (»Kaj je voda?«, »Kdo je Pele?«);
- »Klepetanje« - pogovorne fraze (»Dober dan.«, »Dobro odgovarjaš.«, »Lažeš!«, »Povej kaj pametnega.«, »Ne spreminjaj teme.«);
- »Mašila« - splošni odgovori na neznana vprašanja. S pomočjo tega modula sogovornik sicer ne dobi konkretnega odgovora, vendar je odgovor tako oblikovan, da je vsaj delno v povezavi z vprašanjem. Na vprašanje, ki se začne s »Kako ...?« npr. odgovori »Kako? Ne bi vedel.« in podobno. Tak odgovor je vseeno boljši, kot da bi vedno in na vsa neznana vprašanja odgovarjal enako, npr. »Tega še ne vem.«.

Moduli so jezikovno odvisni in vsebujejo na tisoče vprašanj in odgovorov, ki jih v Amebisu redno nadgrajujemo. Module lahko uporabnik poljubno vklaplja in izklaplja, odvisno od potreb in namena virtualnega pomočnika.

4.2.3. Učenje

Za vnos splošnih odgovorov, ki jih v že vgrajenih modulih ni, ali za vnos specifičnega znanja o podjetju, zaposlenih, izdelkih in storitvah, mora seveda poskrbeti uporabnik sam. V ta namen ima na voljo vrsto možnosti, v razvoju pa so tudi že nekatere nove.

Prva in najenostavnejša možnost učenja je vpis pozdravnega sporočila, s katerim virtualni pomočnik pozdravi ali nagovori sogovornika. Teh je lahko tudi več in se lahko poljubno(krat) spreminjajo. Enako velja za vnos odgovora oz. odgovorov na neznano vprašanje.

Za hiter in enostaven vnos podatkov o nekih tipskih zadevah smo dodali »predloge« (»Templates«). Za podjetja, ki bodo pričakovano najštevilčnejši uporabniki sistema SecondEGO, je mogoče vpisati osnovne podatke o podjetju, njegovih zaposlenih ter izdelkih in storitvah. Prednost predlog je v tem, da je potrebno vpisati le odgovore, pričakovana in možna vprašanja pa so že vgrajena. Glede na potrebe je mogoče izdelati poljubne predloge za poljubne jezike. Predloga »Podjetje« je za zdaj narejena za angleščino in slovenščino.

Za povsem specifične odgovore na specifična vprašanja je mogoče uporabiti t. i. vzorce. Vzorec »ujame« določena vprašanja ali ključne besede in ustrezno odgovori. Del odgovora je lahko tudi prikaz ali zamenjava določene spletne strani, kar dodatno obogati odgovor. Če je vzorcev, ki »ujamejo« vprašanje več, se uporabi odgovor tistega, ki se najbolj ujema s postavljenim vprašanjem. S tem enostavnim mehanizmom lahko vnesemo odgovore na poljubno število vprašanj, pri čemer je vnos hiter in enostaven. Z ustrezno uporabo in povezavo vzorcev je mogoče ustvariti tudi dvosmerno komunikacijo s sogovornikom, kar je lahko zelo uporabno za izvajanje različnih anket ali vodenja sogovornika skozi določen postopek.

Za večjo odzivnost uporabnikov je možna uporaba »sprožilcev« (»Triggers«). Ker uporabniki na spletni strani dostikrat ne opazijo »skritega« virtualnega pomočnika, sprožilci poskrbijo, da se ta sam odpre ob določenih situacijah – po preteku določenega časa, po obisku določenega števila spletnih (pod)strani ali ob obisku točno določene (pod)strani. Sprožilcev je seveda lahko tudi več hkrati, uporabljajo pa se predvsem v marketinške namene (npr. opozarjanje na »akcije«).

Mnogo uporabnikov ima večino informacij, namenjenih svojim (potencialnim) strankam, že na svojih spletnih straneh. Problem nastane, če so te zelo obsežne, in jih stranke ne uspejo (dovolj hitro) najti. V ta namen smo dodali možnost pametnega iskanja po besedilu na spletnih (pod)straneh. Virtualni pomočnik lahko z njegovo pomočjo namesto s konkretnim odgovorom odgovori z naborom povezav na spletne strani, na katerih se najbolj verjetno nahaja odgovor. Na ta način uporabnik najhitreje – brez dodatnega učenja – stranki pomaga do iskanih informacij.

Zahtevnejšim in naprednejšim uporabnikom lahko po dogovoru omogočimo tudi vnos znanja s pomočjo skriptnega jezika, kar je nujno v primeru povezave virtualnega pomočnika z zunanjimi bazami podatkov.

4.2.4. Testiranje

V sistem je vključena možnost testiranja novo vgrajenega znanja, tako da lahko uporabnik vse potencialne napake odpravi še pred objavo.

4.2.5. Analitika

Za uspešno delovanje virtualnega pomočnika je zelo pomembno, da uporabnik sistema sproti preverja, kaj ga ljudje sprašujejo in kako na vprašanja odgovarja. Še posebno na začetku, v prvih tednih in mesecih, da uporabnik pomočnika čim prej nauči odgovorov na najpogostejša vprašanja.

Analitika mu omogoča, da spremlja trenutno, dnevno in mesečno aktivnost (število uporabnikov, število pogovorov, število odgovorov), najpogosteje zastavljena vprašanja v določenem časovnem obdobju, najpogostejša neznana vprašanja in odgovore, uporabo sprožilcev in uporabo vključenih modulov. V razvoju je tudi možnost vpogleda v posamezne pogovore in druge koristne informacije.

Na osnovi statistike vseh vprašanj lahko uporabnik dobi koristne informacije, kaj njegove stranke dejansko zanima. Teh informacij na noben drug način ne more pridobiti, so pa te informacije izredno koristne za izboljšanje ponudbe ali spletnih strani.

Posebno koristna je za uporabnike statistika najpogostejših vprašanj, na katera virtualni pomočnik ne zna odgovoriti, saj so osnova za njegovo učenje.

4.3. Vgradnja na spletno stran

Za uporabo virtualnega pomočnika SecondEGO uporabnik ne potrebuje ne strojne opreme, ne dodatne programske opreme, pač pa le nekaj minut dela. Edina stvar, ki jo mora narediti, je, da v krovno spletno stran doda nekaj vrstic JavaScript kode, ki jo zgenerira sistem SecondEGO.

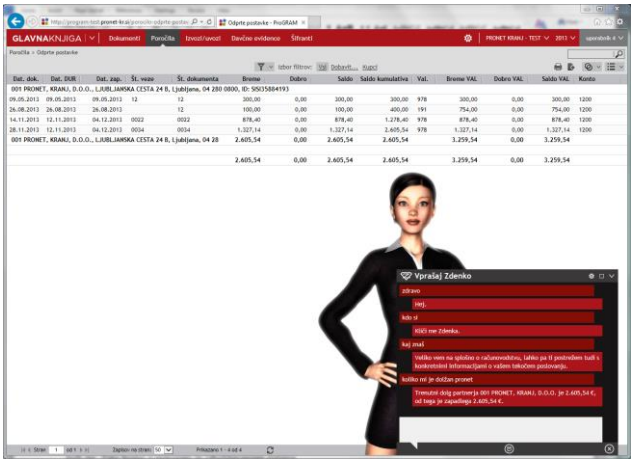
4.4. Vgradnja v druge aplikacije

Virtualnega pomočnika oz. funkcionalnost odgovarjanja na vprašanja v naravnem jeziku je mogoče s pomočjo posebnega vtičnika vgraditi tudi v druge aplikacije. S podjetjem Pronet Kranj smo npr. že izdelali poseben vtičnik za komunikacijo z računovodskimi programi, po naročilu pa je mogoče razviti tudi vtičnike za poljubna področja.

4.5. Delujoči pomočniki

Skupaj s sistemom SecondEGO smo razvijali tudi novo verzijo virtualnega pomočnika Klepca, ki deluje na spletnih straneh podjetja Amebis (www.amebis.si). Namenjen je klepetanju, pa tudi posredovanju informacij o podjetju in njegovih izdelkih.

Poleg Klepca deluje vedno več novih virtualnih pomočnikov SecondEGO. Med bolj uporabljanimi in zanimivimi je virtualna računovodkinja Zdenka (http://program-test.pronet-kr.si/). Vgrajena je neposredno v računovodsko spletno aplikacijo ProGRAM in odgovarja uporabnikom na vprašanja glede konkretnega poslovanja, uporabe programa in splošnih računovodskih zadev. Pri svojem delovanju uporablja podatke iz računovodske baze in zunanje baze besedil s področja računovodstva. Tako je uporabniku – kot v primeru voznega reda iz uvoda – prepuščena odločitev, ali do podatkov pride sam, ali za pomoč zaprosi Zdenko.



Slika7. Virtualna računovodkinja Zdenka

5. Zaključek

Nekaj mesecev delujoč portal SecondEGO že vključuje veliko mehanizmov, s pomočjo katerih je mogoče izdelovati virtualne pomočnike, vendar samo vgrajene možnosti niso dovolj za izdelavo kakovostnih pomočnikov. Za to je s strani uporabnika potrebno vložiti dodatno delo, in več kot ga vloži, boljši je rezultat. Seveda pa lahko z vgradnjo izboljšanih mehanizmov pri tem pomaga tudi Amebis. Zato bomo z razvojem in izboljševanjem sistema glede na želje in zahteve uporabnikov nadaljevali tudi v prihodnje.

6. Literatura

AIML - www.alicebot.org/aiml.html

ChatScript Open Source project – http://sourceforge.net/projects/chatscript/

Wallace, R. (2003). The Elements of AIML Style. A.L.I.C.E. Artificial Intelligence Foundation, Inc.

Wallace, R. (2005) A.L.I.C.E - Artificial Intelligence Foundation http://www.alicebot.org

Weizenbaum, Joseph. "ELIZA - A Computer Program for the Study of Natural Language Communication between Man and Machine," Communications of the Association for Computing Machinery 9 (1966): 36-45.

Ugotavljanje avtorstva besedil: primer »*Trenirkarjev*«

Ana Zwitter Vitez

Oddelek za prevajalstvo, Filozofska Fakulteta
Aškerčeva 2, 1000 Ljubljana
Trojina, zavod za uporabno slovenistiko
Dunajska 116, 1000 Ljubljana
ana.zwitter@guest.arnes.si

Povzetek

V prispevku predstavljamo analizo avtentičnega primera anonimnega besedila, ki je leta 2011 močno vznemirilo slovensko javnost. Avtorstvo besedila smo preverjali na korpusu 75 besedil 21 potencialnih avtorjev na podlagi vnaprej določenega nabora leksikalnih in berljivostnih značilk. Rezultati kažejo, da ima eden od potencialnih avtorjev zelo podobne vrednosti značilk, vendar v dani situaciji ni mogoče preveriti, ali je bil dejanski avtor besedila zajet v analizo ali ne.

Authorship attribution: the 'Sportsuits' case

In this paper we examine an authentic anonymous text which provoked intense reactions in Slovenian media in 2011. Within this authorship attribution task, a corpus of 75 texts written by 21 potential authors was analysed with a predefined set of lexical and readability features. The results show that one of the candidate authors resembles the anonymous text by most of the features although it is not possible to verify whether the actual author was included into the analysis or not.

1. Uvod

Ugotavljanje avtorstva besedil je v zadnjih desetletjih doživelo velik razmah zaradi hitrega razvoja postopkov analize velikih količin besedil, dodatno pa ga spodbuja veliko povpraševanje na področjih prava, kriminologije, literarnih ved in trženja. V zadnjih letih se namreč pogosto soočamo z naslednjimi pojavi:

- plagiati (doktorat nemškega obrambnega ministra, magisterij direktorja Lekarne Ljubljana),
- grozilna pisma (Janez Janša, Katarina Kresal),
- literarni psevdonimi (angleški blog *Belle de Jour*, slovenski roman *Čudoviti klon*),
- analiza profilov strank za potrebe trženja.

Ker je ugotavljanje avtorstva besedil izrazito interdisciplinarno področje, so med obstoječimi pristopi ogromne razlike. Na področju informatike prevladujejo študije z velikimi in dobro označenimi bazami podatkov, kontroliranim naborom značilk in natančno evalvacijo končnih modelov (Sebastiani 2002). Forenzični izvedenci pa se pogosto soočajo s kratkimi besedili brez možnosti analize primerljivega gradiva (Coulthard 2005).

V prispevku predstavljamo avtentičen primer ugotavljanja avtorstva besedila s pomočjo povprečne absolutne razlike med vrednostmi značilk.

2. Ugotavljanje avtorstva besedil

Če je tipična naloga ugotavljanja avtorstva besedil pripisati besedilo neznanega izvora enemu od potencialnih avtorjev, lahko z vidika strojnega učenja isto nalogo opišemo kot klasifikacijo besedil v več razredov (Sebastiani 2002, Stamatatos idr. 2001, Keselj idr. 2003).

Najprej je treba definirati značilke, torej lastnosti besedila, relevantne za klasifikacijo. S pomočjo izračunanih značilk je mogoče določeno besedilo predstaviti v obliki vektorja in tako kvantificirati določene lastnosti besedila. Pri tem različni raziskovalci upoštevajo različne kriterije: leksikalne (Sebastiani 2002, Argamon, Levitan 2005), grafemske (Keselj idr. 2003, Stamatatos 2006), skladenjske (Baayen idr. 1996, Hirst, Feiguina 2007) in semantične (McCarthy idr. 2006).

Na podlagi vektorjev značilk lahko izvedemo klasifikacijo besedil. Nekateri pristopi vsa besedila enega avtorja združijo v en dokument (angl. *compression-based approaches*), nato pa na podlagi te enote poskušajo kvantificirati avtorjev slog (Marton idr. 2005). Drugi pristopi vsako besedilo obravnavajo kot samostojno enoto, ki s svojimi lastnostmi prispeva h gradnji klasifikacijskega modela (Chaski 2005). Pri tem se je kot eden najbolj zanesljivih klasifikatorjev izkazala metoda podpornih vektorjev (SVM), ki ni občutljiva na šum in razpršenost podatkov (Li idr. 2006).

Zadnja etapa ugotavljanja avtorstva besedil je evalvacija. Pri določanju stopnje natančnosti modelov igrajo pomembno vlogo dolžina besedil učnega korpusa (Marton idr. 2005, Hirst, Feiguina 2007), število potencialnih avtorjev (Koppel idr. 2006), in razporeditev besedil na posameznega avtorja (Stamatatos 2008).

3. Kontekst in hipoteza raziskave

Analizirali smo besedilo, ki je leta 2011 močno vznemirilo slovensko javnost. Besedilo je bilo objavljeno na uradni spletni strani ene od parlamentarnih strank in podpisano s psevdonimom Tomaž Majer. Nekaj dni po objavi je informacijska pooblaščenka anonimnega avtorja ovadila zaradi sovražnega govora, sodišče je ovadbo zavrglo, javnost pa se ni nehala spraševati o dejanskem avtorju besedila. Največkrat citirani elementi spornega besedila so se nanašali na interpretacijo zmage nasprotne stranke, ki naj bi ji botrovala udeležba "volivcev s tujim naglasom" in "volivcev v športnih oblačilih (trenirkah), ki so imeli na roki s kemičnim svinčnikom napisano številko, ki jo morajo obkrožiti na glasovnici" Zato se je besedila prijelo ime *Trenirkarji*.

Za potrebe raziskave smo formulirali naslednjo hipotezo: če je avtor besedilo anonimno objavil na uradni spletni strani stranke, obstaja velika verjetnost, da je na isti spletni strani objavil še kakšno svoje besedilo pod drugim ali pravim imenom.
Zato smo za potrebe raziskave analizirali besedila avtorjev, ki so na isti spletni strani objavljali tri mesece pred in tri mesece po objavi spornega besedila. Tako smo dobili korpus 75 besedil 21 podpisanih avtorjev s približno 55.000 pojavnicami, ki so precej neizenačeno razporejene po avtorjih (od 650 do 9000 pojavnic na avtorja).

4. Metodologija

Priprava besedil je zajemala:
- čiščenje besedil,
- pretvorbo iz html v format .txt,
- anonimizacijo besedil (avtorji so označeni z velikimi črkami, njihova besedila pa z zaporednim številom),
- tvorjenje glav dokumentov,[1]
- oblikoslovno označevanje (Grčar idr. 2012) in
- skladenjsko razčlenjevanje besedil (Dobrovoljc idr. 2012).

Analiza besedil je bila zasnovana na vnaprej pripravljenem naboru značilk besedišča in berljivosti[2], s katerim smo se želeli izogniti odvisnosti od tematike besedil.
Leksikalne značilke:
- raznolikost besedišča (*lexical density*),
- Brunetova formula (Brunet 1988): raznolikost besedišča neodvisno od dolžine besedila,
- hapax legomena (Holmes 1992): leme, ki se pojavijo samo enkrat v besedilu,
- Honoréjeva formula (Honoré 1979): razmerje med številom hapaksov in raznolikostjo besedišča.

Berljivostne značilke :
- formula Flesch-Kincaid: razmerje med številom besed in številom povedi
- formula Coleman-Liau: razmerje med številom znakov in številom besed,
- formula Automated Readability Index: izračun stopnje izobrazbe, potrebne za razumevanje besedila ob prvem branju,
- formula Gunning Fog (Gunning 1952): izračun števila let formalnega izobraževanja, potrebnih za razumevanje besedila po prvem branju.

Na podlagi naštetih formul smo izračunali povprečne absolutne razlike med vrednostmi značilk in postavili hipotezo o najverjetnejšem avtorju anonimnega besedila.

5. Analiza

[1] Primer anonimizirane glave besedila: *A_1: Politična izprijenost in pošteni civilisti*
[2] Formule: **Lexical Density** = (different words / words) x 100, **Gunning Fog Index** = 0.4 x (ASL + ((SYW / words) x 100)), **ARI** = (0.5 x ASL) + (4.71 x ALW) - 21.43 **Coleman-Liau Grade** = 5.89 x ACW - 0.3 x sentences / (100 x words) - 15.8 **Flesch-Kincaid Grade Level** = (0.39 x ASL) + (11.8 x ASW) – 15.59

Najprej smo izračunali absolutne vrednosti značilk besedišča in berljivosti za vsako od analiziranih besedil. Tabeli 1 in 2 predstavljata rezultate za anonimno besedilo:

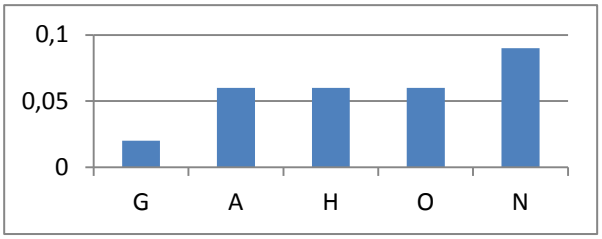
Značilka	**Vrednost**
Raznolikost besedišča	0,38
Formula Brunet	12,96
Statistika Honoré	1998,79
Hapax legomena	0,24

Tabela 1: Leksikalne značilke anonimnega besedila.

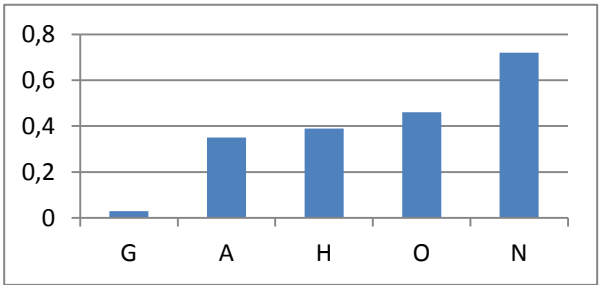
Značilka	**Vrednost**
Razmerje št. besed/št. povedi	21,24
Razmerje št. znakov/št. besed	5,14
Indeks ARI	13,38
Formula Gunning Fog	21,81

Tabela 2: Berljivostne značilke anonimnega besedila.

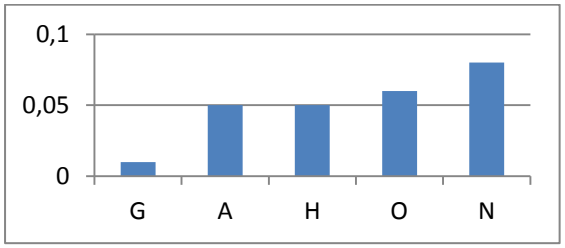
Na podlagi izračunanih formul smo opazovali povprečno absolutno razliko med vrednostmi značilk znanih avtorjev in anonimnega besedila.
V nadaljevanju (grafi 1 do 8) predstavljamo rezultate razvrščanj glede na upoštevane značilke. Pri vsaki značilki predstavljamo prvih pet avtorjev z najmanjšo povprečno absolutno razliko glede na anonimno besedilo (najmanjša povprečna absolutna razlika pomeni največjo podobnost z anonimnim besedilom).



Graf 1. Besedišče.
Graf 1 prikazuje razvrstitev avtorjev glede na razmerje med številom različnih lem in celokupnim številom lem v besedilu. Najbliže anonimnemu besedilu so vrednosti avtorja G.

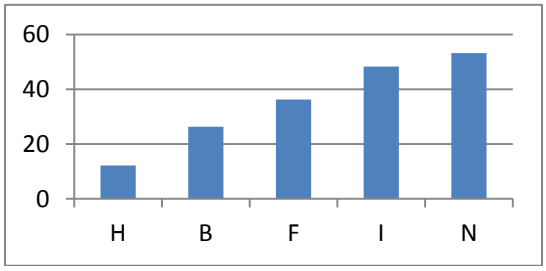


Graf 2. Brunet.
Graf 2 predstavlja avtorje z najmanjšo povprečno absolutno razliko glede na Brunetovo formulo, ki računa raznolikost besedišča glede na dolžino besedila.



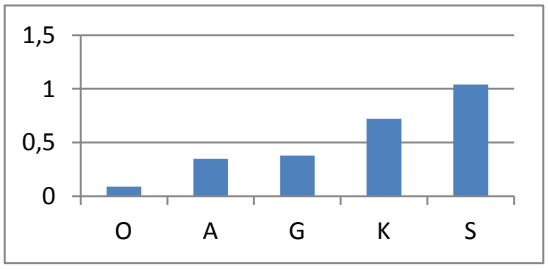
Graf 3. Hapax legomena

Graf 3 predstavlja avtorje, razvrščene glede na število lem, ki se pojavijo samo enkrat v besedilu. Tudi po tem kriteriju se avtor G najmanj razlikuje od anonimnega besedila.



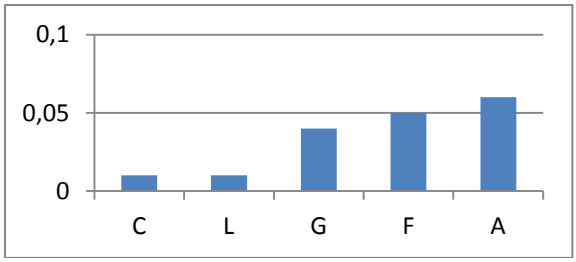
Graf 4. Honoré.

Graf 4 razvršča po formuli Honoré (Honoré 1979), ki računa razmerje med številom hapaksov in raznolikostjo besedišča. Najmanjšo povprečno absolutno razliko ima avtor H.



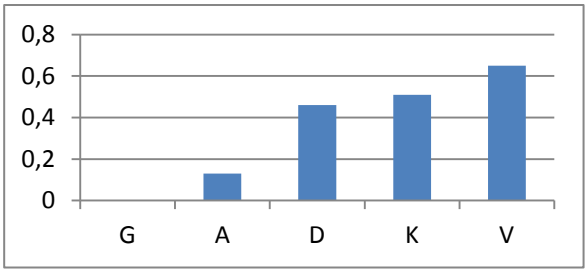
Graf 5. Število besed/število povedi.

Razmerje med številom besed in številom povedi, ki ga prikazuje graf 5, je pogosto uporabljan kriterij za določanje stopnje berljivosti besedila.



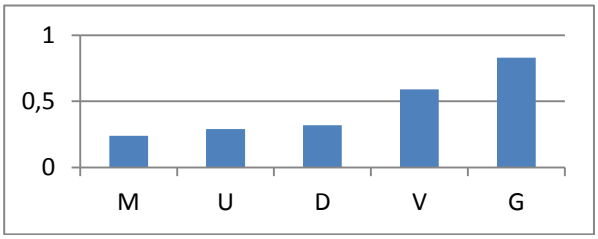
Graf 6. Število znakov/število besed.

Graf 6 izpostavlja razmerje med številom znakov in številom besed v analiziranih besedilih. Po tem kriteriju ima najmanjšo povprečno absolutno razliko od anonimnega besedila avtor C.



Graf 7. ARI.

Formula ARI (Automated Readability Index) izračuna okvirno stopnjo izobrazbe, ki jo zahteva neko besedilo za razumevanje ob prvem branju. Ta značilka izpostavlja avtorja G, ki po tem kriteriju dosega identične vrednosti kot anonimno besedilo.

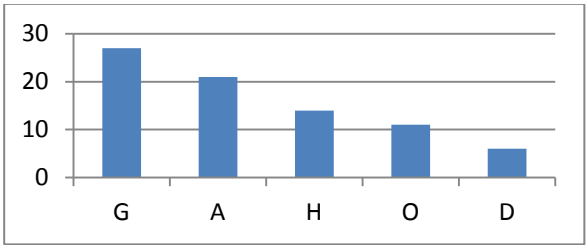


Graf 8. Gunning Fog.

Formula Gunning Fog predstavlja nekoliko drugačen izračun stopnje izobrazbe, ki jo zahteva neko besedilo, da ga bralec razume po prvem branju. Glede na ta kriterij kaže največjo podobnost z anonimnim besedilom avtor M.

6. Sinteza in omejitve raziskave

Če vsakemu od prvih petih avtorjev z najmanjšo povprečno absolutno razliko glede na anonimno besedilo pripišemo od 1 do 5 točk, dobimo naslednjo razvrstitev:



Graf 9: Stopnja podobnosti avtorjev z anonimnim besedilom.

Raziskava je nastala v okviru avtentične situacije objave anonimnega besedila, ki je vznemirila slovensko javnost. Za potrebe raziskave smo postavili hipotezo, da je avtor na isti spletni strani verjetno objavil še kakšno drugo besedilo in ga podpisal s svojim pravim imenom. Hipoteza je verjetno utemeljena, vendar raziskava:

- ponuja premajhen nabor besedil, da bi lahko izvedli evalvacijo modela (Argamon, Levitan 2005),

- ne ponudi odgovora na vprašanje, ali je bil dejanski avtor besedila sploh vključen v analizo.

7. Zaključek

V prispevku poskušamo identificirati avtorja anonimnega besedila, ki je v slovenskih medijih sprožil številne odzive. Analiza zajema 75 besedil 21 znanih avtorjev in vključuje razvrščanje na podlagi leksikalnih in berljivostnih značilk.

Rezultati kažejo, da je med 21 potencialnimi avtorji glede na upoštevane kriterije razvrščanja najverjetnejši avtor anonimnega besedila avtor G. Vendar zaradi majhnega števila analiziranih besedil ne moremo izvesti evalvacije razvrščanja in preverjanja hipoteze.

Analiza bi pridobila na kredibilnosti, če bi lahko analizirali večje število besedil, potrdili razlikovalno moč upoštevanih značilk na večji bazi besedil ali dopolnili metodo s kvalitativno analizo diskurza.

Primarna vira

Besedilo *Volivci v trenirkah*:
http://www.delo.si/assets/media/other/20111211//Prispevek%20Toma%C5%BEa%20Majerja.pdf
Spletni arhiv stranke:
http://www.sds.si/arhiv?id=12

Bibliografija

S. Argamon, S. Levitan. 2005. Measuring the usefulness of function words for authorship attribution. P*roceedings of ACH/ALLC 2005*.

attribution. In Proceedings of the Pacific

R. Baayen, H. van Halteren, F. Tweedie. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11/3, str. 121–131.

E. Brunet. 1988. Une mesure de la distance intertextuelle : la connexion lexicale. *Le nombre et le texte. Revue informatique et statistique dans les sciences humaines*, 24/1, str. 81–116.

C. E. Chaski. 2005. Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4/1, str. 1–14.

M. Coulthard. 2005. The linguist as expert witness. *Linguistics & the Human Sciences*, 1/1, str. 39–58.

K. Dobrovoljc, S. Krek, J. Rupnik. 2012. Skladenjski razčlenjevalnik za slovenščino. *Zbornik Osme konference Jezikovne tehnologije*, str. 42-47.

M. Grčar S. Krek, K. Dobrovoljc. 2012. Obeliks: statistical morphosyntactic tagger and lemmatizer for Slovene. *Zbornik Osme konference Jezikovne tehnologije*, str. 42–47.

R. Gunning. 1952. *The Technique of Clear Writing*. New York: McGraw-Hill.

G. Hirst, O. Feiguina. 2007. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22/4, str. 405– 417.

D. I. Holmes. 1992. A Stylometric Analysis of Mormon Scripture and Related Texts. *Journal of the Royal Statistical Society*, 155/1. str. 91-120.

A. Honoré. 1979. Some Simple Measures of Richness of Vocabulary. *Association of Literary and Linguistic Computing Bulletin*, 7, str. 172–177.

V. Keselj, F. Peng, N. Cercone, C. Thomas. 2003.. N-gram-based author profiles for authorship attribution. *Proceedings of the Pacific Association for Computational Linguistics*, str. 255-264.

M. Koppel, J. Schler, S. Argamon, E. Messeri. 2006. Authorship attribution with thousands of candidate authors. *Proceedings of the 29th ACM SIGIR*, str. 659-660.

K. Luyckx, W. Daelemans. 2005. Shallow text analysis and machine learning for authorship attribution. *Proceedings of the Fifteenth Meeting of Computational Linguistics in the Netherlands*. str. 149–160.

Y. Marton, N. Wu, L. Hellerstein. 2005. On compression-based text classification. *Proceedings of the European Conference on Information Retrieval*, str. 300–314.

P. M. McCarthy, G. A. Lewis, D. F. Dufty, D. S. McNamara. 2006. Analyzing writing styles with coh-metrix. *Proceedings of the Florida Artificial Intelligence Research Society International Conference*, 764-769.

F. Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*,34/1, str. 1-47

E. Stamatatos. 2006. Authorship attribution based on feature set subspacing ensembles. *International Journal on Artificial Intelligence Tools*, 15/5, str. 823-838.

E. Stamatatos. 2008. Author identification: Using text sampling to handle the class imbalance problem. Information Processing and Management, 44/2, str. 790-799.

E. Stamatatos, N. Fakotakis, G. Kokkinakis. 2001. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35/2, str. 193-214.

E. Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*. 60/3, str. 538–556.

R. Zheng, J. Li, H. Chen, Z. Huang. 2006. A framework for authorship identification of online messages: Writing style features and classification techniques. *Journal of the American Society of Information Science and Technology*, 57/3, str. 378-393.

A. Zwitter Vitez. 2012. Authorship Attribution: Specifics for Slovene. *Slavia Centralis* 5/1, str. 75–85.

Razreševanje sklicev pri analizi slovenskih besedil

Peter Holozan

Amebis, d. o. o.
Bakovnik 3, 1241 Kamnik
peter.holozan@amebis.si

Povzetek

Razreševanje sklicev je pomemben del jezikovnih tehnologij, vendar za slovenščino ta tehnologija še ni bila razvita. Obstajajo različne vrste sklicev, članek se osredotoča predvsem na anafore pri osebnih zaimkih. Uporabljene so bile štiri metode razreševanja, ki se med seboj dopolnjujejo, najpomembnejša temelji na metodah na osnovi aktivacije. Prvi rezultati so obetavni, razreševanje sklicev je bilo uporabljeno tudi v sistemu za odgovarjanje na vprašanja Piflar, ki zna s tem odgovoriti na več vprašanj.

Reference Resolution for Slovenian Texts Analysis

Reference resolution is an important part of language technologies, but has not yet been developed for Slovenian. There are various types of references and the paper focuses on anaphora resolution of personal pronouns. Four methods, used in combination, were used; the most important one is based on activation. First results are promising: reference resolution was used in the question answering system Crammer, which can, as a result, answer more questions than before.

1. Uvod

Razreševanje sklicev je pomemben del jezikovnih tehnologij. Sklice lahko razdelimo na anafore, kjer je razlaga pred sklicem, in katafore, kjer je razlaga za sklicem. Tipičen primer sklicev so zaimki, niso pa sklici omejeni le na zaimke, čeprav se največkrat omejimo le nanje.

Razreševanje sklicev lahko pomaga tudi pri razdvoumljanju besedil, saj v precej primerih tega ne moremo narediti brez razreševanja sklicev, iz česar sledi, da v resnici ne moremo izhajati iz tega, da imam vhodno besedilo že razdvoumljeno, ampak se morata razdvoumljanje in razreševanje sklicev dopolnjevati. (McShane, Beale, Nirenburg, 2010) Tak primer sta npr. povedi »Miha je videl matico, ki jo je privil Janez.« in »Miha je videl matico, ki jo je vzgojil Janez..«, kjer je pomen besede »matica« odvisen od prilastkovega odvisnika, kjer je »matica« nadomeščena z osebnim zaimkom »jo«.

V veliko primerih se razreševanje omeji le na razreševanje anafor. (Mitkov, 1999; Němčík, 2006)

Za slovenščino razreševanje sklicev še ni bilo narejeno, zato je smiselno preizkusiti, kako uspešno se to da vgraditi v analizator, ki prevaja naravni jezik v Amebisov vmesni jezik, katerega podrobni opis je v prilogi 6.2 v (Holozan, 2011). Ta vmesni jezik uporabljajo mnogi izdelki podjetja Amebis, npr. strojni prevajalnik Presis in sistem za odgovarjanje Piflar, kar pomeni, da bo ta izboljšava vplivala tudi nanje.

Najprej bo v razdelku 2 predstavljen problem sklicev, pri čemer bodo omenjene razlike pri sklicih med slovenščino in angleščino, za katero je bilo opravljeno največ dela pri razreševanju sklicev.

V razdelku 3 bodo opisane nekatere obstoječe metode razreševanja sklicev.

Nato bodo v razdelku 4 preizkušene različne metode za razreševanje sklicev, ki se izvajajo zaporedno in iz katerih se dobi skupni rezultat označevanja referenc.

Na koncu bodo v razdelku 4 predstavljeni rezultati, v razdelku 5 pa bo razreševanje sklicev uporabljeno v sistemu za odgovarjanje na vprašanja v naravnem jeziku Piflar.

2. Sklici

Glavni kandidati za sklice so zaimki, in sicer predvsem osebni. V Gigafidi pokrivajo osebni zaimki 1,4 % vseh besednih pojavnic, vendar je treba upoštevati še, da v slovenščini velik del osebnih zaimkov izpuščamo (v povedih iz korpusa jos100k (Erjavec, Krek, 2008), ki jih je analizatorju uspelo analizirati, je izpuščenih osebnih zaimkov več kot dvakrat toliko kot neizpuščenih), jih je pa vseeno treba razrešiti, če želimo dobiti pravi pomen teh stavkov.

V avtodomu sta ***onadva*** prevažala opij

Policisti in kriminalisti so v sodelovanju s cariniki na avtocestnem počivališču v bližini Murske Sobote zaustavili tovorni avtomobil, ki je bil predelan v avtodom. Po pregledu avtomobila so *oni* v njem našli 390 gramov opija, tri vrečke obrezanih makovih glavic in tri grame halucinogenih gobic. Vozilo je bilo registrirano v Franciji. V njem pa sta se vozila **Francoza**, stara 32 in 33 let, so sporočili *oni* s policijske uprave v Murski Soboti.

Zaradi utemeljenega suma neupravičene proizvodnje in prometa s prepovedanimi drogami sta bila **tujca** s kazensko ovadbo privedena pred preiskovalnega sodnika okrožnega sodišča v Murski Soboti, ki je zoper **oba** odredil pripor.

Slika 1. Primer.

Slika 1 prikazuje primer besedila s sklici. V kurzivi so dodani sicer izpuščeni osebni zaimki, odebeljeno pa so označeni sklici, ki jih je treba povezati. Podobno lahko povežemo še »avtodomu«, »tovorni avtomobil«, »ki«, »avtomobila«, »vozilo« in »njem«.

Razreševanje sklicev, ki niso zaimki, je v slovenščini težje kot v angleščini, ker slovenščina ne uporablja členov, zato se iz tega, da ima neka samostalniška fraza določni člen, ne da sklepati na to, da se nanaša na nekaj, kar je bilo omenjeno že prej, kot je to primer v angleščini. To pomeni, da se moramo v slovenščini tukaj veliko bolj

zanesti na pomene, delno pa tudi na besedni red (členjenje po aktualnosti).

Težava so tudi sklici, ki so v angleščini zapisani z besedo »one«, npr.: »If you cannot attend a tutorial in the morning, you can go for an afternoon one.« (Mitkov, 1999). Slovenski prevod bi bil: »Če se ne moreš udeležiti vaj dopoldne, greš lahko na popoldanske.« V slovenščini tukaj ni posebne besede, na katero bi lahko vezali sklic, ampak bi tukaj lahko rekli, da gre za izpust besede »vaje« v drugem stavku.

Sklici lahko povezujejo več predhodnih besed v eno besedo ali obratno. V primeru »Srečal sem Johna in Mary. Bila sta zelo vesela, saj smo dobri prijatelji.« se John in Mary najprej povežeta v izpuščeni zaimek »onadva«, nato pa še skupaj s pripovedovalcem (1. osebo) v izpuščeni zaimek »mi«. Obratno pa je v primeru »Starejši par je hodil po parku in moški se je nenadoma spotaknil.«, ko je »moški« najverjetneje del »para« iz predhodnega stavka.

Razreševanje sklicev je zelo odvisno od pomenov. Če uporabim primer iz (Němčík, 2006): »John je skril Billove ključe. Bil je pijan.«, se ljudem zdi najverjetnejša interpretacija, da se drugi stavek nanaša na Billa, ker pač sklepamo, da je vožnja pod vplivom alkohola nevarna in je Johna skrbelo za Billa, zato mu je skril ključe, da ne bi mogel odpeljati. Ni pa to edina možna interpretacija, morda je John bil pijan in je hotel nagajati Billu in mu je zato skril ključe hiše. Taki primeri kažejo na to, da je razreševanje sklicev res zahteven problem za računalnike.

3. Nekatere metode razreševanja sklicev

Za razreševanje sklicev je bila razvita množica metod in nekatere bodo na kratko predstavljene v nadaljevanju tega razdelka.

3.1. Hobbsovo sintaktično iskanje

Hobbsovo sintaktično iskanje (Hobbs, 1978) je bila prva metoda, ki je uporabila jezikovno znanje in je kljub starosti in relativni preprostosti (že sam Hobbs je menil, da je to le naivna metoda) še vedno primerljivo uspešna v primerjavi z modernejšimi metodami (Němčík, 2006).

Osnova za postopek je drevo izpeljav za poved. Hobbsovo iskanje določi vrstni red, v katerem samostalniške fraze postanejo kandidate za razreševanje sklicev. V drevesu začnemo iskati levo od zaimka, za katerega želimo razrešiti sklic, potem pa se dvigamo in vsakič iščemo v širino od leve proti desni.

Metodo lahko dopolnimo s pomenskimi omejitvami pri kandidatih.

Težava pri metodi je, da lahko vedno najdemo primere, v katerih ne deluje, dodatno pa je izdelava drevesa izpeljav sama po sebi zapleten problem.

3.2. Algoritem BFP

Algoritem BFP (Brennan, Friedman, Pollard, 1987) temelji na teorija fokusa (centering theory), ki je bila prvič opisana v (Joshi, Kuhn, 1979). Ta opisuje, kako se spreminja fokus diskurza, ena od metod fokusiranja pa je tudi uporaba zaimkov, ki nas usmerjajo na fokus. Ta se lahko spreminja z različnimi vrstami prehodov.

Pokazalo pa je se, da razvoj v smeri vedno bolj kompleksnih pravil slepa ulica, ker ni bilo mogoče dovolj podrobno zajeti splošnega znanja in opisati jezika, zato so se metode usmerile v smeri, ki zahteva manj znanja (Němčík, 2006).

3.3. Faktorji poudarka

Postopek s faktorji poudarka (salience factors) je bil predlagan v (Lappin, Leass, 1994). Ti faktorji so uteži, ki so prirejene posamičnim možnostim sklicev in potem kombinirane, da se določi najpomembnejši element diskurza. Dodatno postopek ugotavlja, kateri zaimki so del fraz in nimajo sklicev (npr. »it« v »It's raining.«) in določa povratne zaimke (Němčík, 2006).

Uteži je treba določiti z eksperimentiranjem, kar pomeni, da potrebujemo korpus primerov, da lahko avtomatsko preverjamo različne uteži.

3.4. Robustni sistemi z malo potrebnega znanja

Primer za tak sistem je MARS (Mitkov, Evans, Orasan, 2002). Sistem temelji na množici predhodnostnih kazalnikov (set of antecedent indicators). Vsak od njih opisuje določen pogoj, ki se nanaša na danega kandidata za sklic, in vpliv, ki ga ima na verjetnost, da je to verjetni izvor sklica (Němčík, 2006).

Prednost te metode je, da ne potrebuje zunanjega skladenjskega razčlenjevalnika, za večino kazalnikov pa se zdi, da je jezikovno neodvisna, zato je bila te metoda uporabljena tudi za druge jezike, kot so francoščina, poljščina, arabščina in bolgarščina (Němčík, 2006).

3.5. Statistične metode

Po letu 1990 so se za razreševanje sklicev začele uporabljati tudi statistične metode (in tudi druge metode strojnega učenja). Primer je (Ge, Hale, Charniak, 1998).

Vendar vse te metode zahteva korpus učnih primerov, ki ga za slovenščino še nimamo, zato se za zdaj nismo usmerili v to smer.

4. Uporabljene metode razreševanja

Ideja razreševanja sklicev, ki jo opisujemo v nadaljevanju, je uporaba množice metod, od katerih vsaka razrešuje določene sklice, metode pa se uporabljajo od bolj proti manj zanesljivim (v tem vrstnem redu so tudi opisane, poudariti pa je treba, da je zanesljivost v tem trenutku le ocena, ki potrebuje še bolj temeljito preverjanje na večjem številu primerov).

Izbrane so bile metode, ki ne potrebujejo učnega korpusa, ker tega za slovenščino še ni. Obstaja pa po drugi strani možnost, da bi si lahko s temi za zdaj uporabljenimi metodami pomagali, da se naredi osnutek korpusa primerov sklicev, ki se potem še ročno dopolni, da ni treba celotnega izdelati ročno.

Sklici so v vmesnem jeziku opisani z novim elementom ORI, ki je dodan k obstoječemu elementu (največkrat je to osebni zaimek (OSZ) oz. navidezni (ki se skriva v osebni glagolski obliki) osebni zaimek (NOZ), lahko pa tudi drug samostalniški zaimek (SAZ) ali samostalnik (SAM)) v element JED (jedro dela samostalnike fraze) in element ORI vsebuje element SFR (samostalniško frazo). Slika 2 prikazuje primer, ko je sklic dodan osebnemu zaimku v slogi osebka (element OSB).

(1OSB:(-SFR:(-DSF:(-JED:(-OSZnemt:[10])**,(-ORI:(-SFR:(-DSF:(-JED:(-SAMe:{7d62a7;4207ac9}[/]<dc>)))))**)))))

Slika 2. Primer zapisa sklica v vmesnem jeziku

Za preizkušanje so bili uporabljeni umetno skonstruirani primeri, pravljica Rdeča kapica, Cankarjev Na klancu, testno besedilo iz priročnika Pravipis Aleksandre Kocmut, Wikipedija, šala neznanega izvora in prispevek iz črne kronike.

4.1. Izpusti osebka

To je vrsta sklicev, ki jih je mogoče zelo zanesljivo razrešiti. Gre za zaporedna stavka, pri čemer je v drugem izpuščen osebek, tako da se uporabi kar osebek iz prvega stavka: »Miha je prišel do vrat in pozvonil.«. V teh primerih se običajno izpusti še pomožni glagol, lahko pa tudi veznik: »Metka je rekla, da rada pleše in poje.«

4.2. Prilastkovi odvisniki

Tudi pri prilastkovih odvisnikih vemo, da se zaimek (»ki«, »kateri« ali pa naslonska oblika osebnega zaimka ob »ki«) v odvisniku nanaša na besedo, ki je jedro ob tem odvisniku. V primeru »Miha je videl sliko, ki jo je naslikal Janez.« tako vemo, da se »jo« nanaša na besedo »sliko«.

Težava lahko nastopi le v primerih, ko ni jasno, kaj je jedro: »Bila sta privedena pred preiskovalnega sodnika okrožnega sodišča v Kamniku, ki je zoper oba odredil pripor.« V takih primerih se lahko zgodi, da analizator označi kot jedro »Kamnik«, kar morajo potem razrešiti pomenske omejitve.

4.3. Delna osebna imena

Še posebej v časopisnem poročanju je običajno, da se oseba prvič navede s polnim imenom, v nadaljevanju pa le s priimkom (v bolj neformalnih besedilih pa tudi le z imenom), npr.: »*Darko Krašovec* je bil ponoči, na prvi seji pravkar oblikovane vlade Mira Cerarja, potrjen za generalnega sekretarja. Čeprav do zdaj sodnik, pa **Krašovec** v politiki ni novo ime.« Pri časopisnih naslovih so pogoste tudi katafore take vrste, saj je oseba v naslovu omenjena le s priimkom, v samem članku pa je potem navedena s polnim imenom.

Postopek za to vrsto sklicev pravzaprav ni posebej zapleten, če imamo podatek, kaj so osebna imena, vsa imena oseb je treba shraniti v seznam in potem pogledati po seznamu, kadar naletimo le na posamičen priimek oz. ime.

Zapis, ki ga uporablja Amebisov vmesni jezik, ki prvi del imena osebe (običajno torej osebno ime) zapiše v elementu JED (jedro dela samostalniške fraze), priimek pa v elementu IMP (imenski prilastek), po drugi strani pa sam priimek postane JED (če pa je pred imenom še kakšna druga beseda, npr. »matematik Josip Plemelj«, pa celo tako osebno ime kot priimek postaneta IMP), sicer pomeni, da se postopek malo zaplete in je treba pri izvedbi paziti na vse te pretvorbe. Dodatna težava so primeri, kjer bi morali sklic vezati na element IMP, kar za zdaj še ni podprto (če je torej posamičen priimek uporabljen kot prilastek za drugo besedo, npr. »matematik Plemelj« kot sklic za »matematik Josip Plemelj«).

4.4. Anafore pri osebnih zaimkih

Postopek za razreševanje anafor je bil zasnovan na podlagi metod na osnovi aktivacije (activation-based methods), kakor so opisane v (Němčík, 2006) in ki izhajajo iz dela Eve Hajičove in sodelavcev, vendar v tem trenutku še v precej poenostavljeni in predelani obliki.

Postopek je tak, da se gradi kontekst analize, ki vsebuje seznam kandidatov za razreševanje anafor, pri čemer ima vsak kandidat shranjeno analizo ustrezne samostalniške fraze, mesto zadnje uporabe (npr. osebek, predmet v tožilniku, prislovno določilo), podatke o spolu, številu, osebi in živosti ter oceno. Ko se pride do osebnega zaimka, ki še nima razrešenega sklica, se poišče, ali obstaja kakšen kandidat, ki ustreza glede spola, števila, osebe in živosti, če jih je več, se izbere tisti, ki ima višjo oceno oz. se je pojavil zadnji, dodatno pa oceno zviša še ujemanje mesta uporabe (če npr. razrešujemo sklic pri osebku, ima prednost kandidat, ki je bil že prej osebek).

Uporaba kandidata mu poviša oceno, z začetno oceno se na seznam kandidatov dodajo tudi vse samostalniške fraze, ki nastopajo v analizi. Na koncu vsakega stavka, povedi in odstavka se znižajo (prepolovijo) ocene vseh obstoječih kandidatov, kandidati, katerih ocena pade na 0, se izbrišejo iz konteksta analize.

Ta osnovni postopek je bil dopolnjen z dodatnimi pravili, ki so opisana v nadaljevanju.

4.4.1. Premi govor

Premi govor prekine tok pripovedovanja z drugim tokom, zato konteksta iz spremnega besedila ne smemo uporabiti pri analizi premega govora in obratno. Rešitev je, da ima analizator dva konteksta – enega za osnovno besedilo in drugega za premi govor, pri čemer se kontekst za premi govor vsakič ponastavi (dokler ne bo izdelana boljša analiza diskurza, ki bi določila, kdo se s kom pogovarja).

Dopolnitev za prihodnost je še, da se iz spremnega stavka v kontekst premega govora preneseta prva in druga oseba (iz »Janez je rekel Micki: 'Jutri ti bom prinesel to knjigo.'« bi tako lahko ugotovili, da bo Janez prinesel knjigo Micki).

4.4.2. Pomenske omejitve

Samo informacije o skladnji in osnovne omejitve (oseba, spol, število) ne zadoščajo vedno za razreševanje sklicev.

Metka je prebrala knjigo, ki jo je napisala Karmen, in jo povabila na kavo.
Metka je prebrala knjigo, ki jo je napisala Karmen, in jo vrgla stran.

Slika 3. Pomenske omejitve pri sklicih

Čeprav sta si povedi na sliki 3 enaki do drugega »jo«, je razrešitev tega sklica vseeno drugačna. V prvi povedi se drugi »jo« nanaša na »Karmen«, v drugi pa na »knjiga«.

Podobno je v realnem primeru »Zadremala je že skoro, ali zgodilo se ji je, kakor da bi polagoma drsala navzdol, kakor da bi se skrinja nagibala, nagibala ... in prestrašila se je in se je prebudila.«, kjer se ni prestrašila skrinja, ampak oseba, ki sicer ni navedena v tej povedi.

V precejšnjem delu primerov si bo dalo pomagati že s tem, da imajo glagolske predloge lahko pri parametrih omejitve, ali so ti parametri obvezno osebe (oz. organizacije) oz. niso osebe. Vendar pa to vedno ne zadošča, v primeru »Hm, lahko bi kar takoj pojedel to deklico, ampak je premajhna, da bi mi potešila lakoto. Če

odigram pravilno, bom lahko pojedel njo, pa tudi njeno babico!« je tako postopek najprej menil, da se »njo« nanaša na »lakoto« kar pomeni, da je treba v predlogi omejiti, da se ne da pojesti lakote. V takih primerih bi si lahko pomagali tudi s korpusom, vendar oseb ne jemo prav pogosto, če ne gre za pravljico.

4.4.3. Stavki brez analize

Pojavi se vprašanje, kaj narediti v primeru, ko analizatorju ne uspe analizirati katerega od vmesnih stavkov. Tak primer je bil »Sončni žarki so se že igrali na strehi županove hiše. Francka je bila vsa nemirna, srce ji je utripalo od sreče in obenem od straha, da bi zamudila voz.«, kjer analizator ni prepoznal stavka »Francka je bila vsa nemirna« (ker še ni podpiral kombinacije zaimka »ves« s pridevnikom na mestu povedkovega določila), zato je potem postopek priredil zaimku »ji« vrednost »županova hiša«.

Idealna rešitev je, da se dopolni analizator, vendar ni mogoče pričakovati, da mu bo v dogledni prihodnosti uspelo analizirati vse (še posebej pri izpustih) zato je varianta, ki je vredna razmisleka, ta, da se v takih primerih ponastavi (pobriše) stanje konteksta. Na ta način sicer lahko izgubimo nekatere razrešitve sklicev, ki izhajajo še iz prejšnjih stavkov, vendar se izognemo napakam, kar je v večini primerov bolj pomembno (torej povečamo natančnost na račun priklica).

Vsekakor pa je dolgoročna rešitev izboljševanje analizatorja.

4.4.4. Ponavljanje v stavku

V primeru »Ko sva zapuščala hišo, se je mačka nekako med nogami izmuznila nazaj v hišo. Nisva jo želela pustiti v hiši, ker se neprestano trudi požreti papigo.« je analizator poskušal razrešiti »jo« s »hiša« namesto »mačka«. Pomenske omejitve ni (morala bi biti precej podrobna, kjer lahko nekje pustiš tako osebo kot predmet), možno pa je postaviti dodatno zahtevo, da ne smemo znotraj posameznega stavka razrešiti zaimka z besedo, ki v tem stavku še nastopa, s čimer se potem zaimek »jo« razreši v »mačka«.

4.4.5. Pomenske razlike med izpuščenimi in navedenimi osebnimi zaimki

Lastnost, ki nam lahko pomaga pri razreševanju sklicev, je, da se na neživo vedno nanašamo le z naslonsko obliko osebnega zaimka. Tako ne moremo reči »Knjiga je bila zelo zanimiva in njo sem prebral v dveh urah.« ampak le »Knjiga je bila zelo zanimiva in prebral sem jo v dveh urah.« Ker v imenovalniku ni naslonskih oblik, to pomeni, da neživo v osebku ne more biti nadomeščeno z osebnim zaimkom, ampak le z izpustom osebnega zaimka ali pa s kazalnim zaimkom.

S pomočjo tega pravila lahko v besedilu »Njegova jeza je v Naomi zbujala občutek krivde. Ona je bila tista, ki je vztrajala na prenovi strehe.« ugotovimo, da »Ona« ne smemo razrešiti z »njegova jeza«, ampak z »Naomi«.

4.5. Prislovni zaimki

Tukaj se razrešujejo sklici, ki so vezane ne prislovne zaimke, in sicer na »tam«, »tja« in »takrat«. Prva dva se sklicujeta na kraj, drugi pa na čas.

Oblaki nastajajo poleti *nad večjimi ognjeniki*. **Tam** nastanejo zato, ker se topli zrak dviga in ohlaja.
Princ Borjatinski, guberner Jakutska je leta 1670 zaupal Dežnjovu odpravo *v Moskvo*. **Tja** je moral odnesti »sobolji zaklad« in uradne dokumente.
Biologija se je začela hitro razvijati in rasti, *ko je Anton van Leeuwenhoek izboljšal mikroskop*. **Takrat** so učenjaki odkrili semenčice, bakterije, infuzorije in raznovrstnost mikroskopskega življenja.
Večina dragocenosti, ki jo jih Slovani naropali v Hersonu, se je znašla v Novgorodu, kjer so jih vse do 20. stoletja hranili *v katedrali sv. Sofije*. **Tja** so prišle morda po zaslugi prvega novgorodskega škofa Joahima Hersonskega, katerega ime kaže na njegovo povezanost s tem mestom.
Prosti čas je mojster izkoristil za obisk *Londona*. **Tja** je prišel kot izrazit skladatelj italijanske opere.

Slika 4. Primeri s prislovnimi zaimki

Slika 4 vsebuje primere prislovnih zaimkov s sklici. V prvih treh primerih zaimek nadomešča prislovno določilo iste vrste (z isto vprašalnico), četrti primer pa kaže, da je pri krajevnih prislovnih določilih treba imeti možnost, da se pretvarja med prislovnimi določili kraja za »kje« in »kam« (torej je treba »v katedrali sv. Sofije« pretvoriti v »v katedralo sv. Sofije«), kar za zdaj v Amebisovem vmesnem jeziku ni možno, zato bo treba ustrezno dopolniti podatkovno bazo Ases s povezavami med pomeni predlogov oz. prislovov.

Zadnji, peti primer pa kaže, da ni nujno, da se prislovni zaimki sklicujejo na prislovna določila, ampak se lahko sklicujejo tudi na samostalnike, npr. na zemljepisna imena.

Razmislek pri prislovnih zaimkih je, da so relativno redki v besedilih, uporabljamo jih le, če želimo povezavo posebej poudariti. Veliko pogostejše je implicitno navezovanje, da se naslednji stavek dogaja v istem času in prostoru, zato bo treba razmišljati tudi v smeri, kako najti te implicitne sklice.

4.6. Katafora v osebku odvisnika

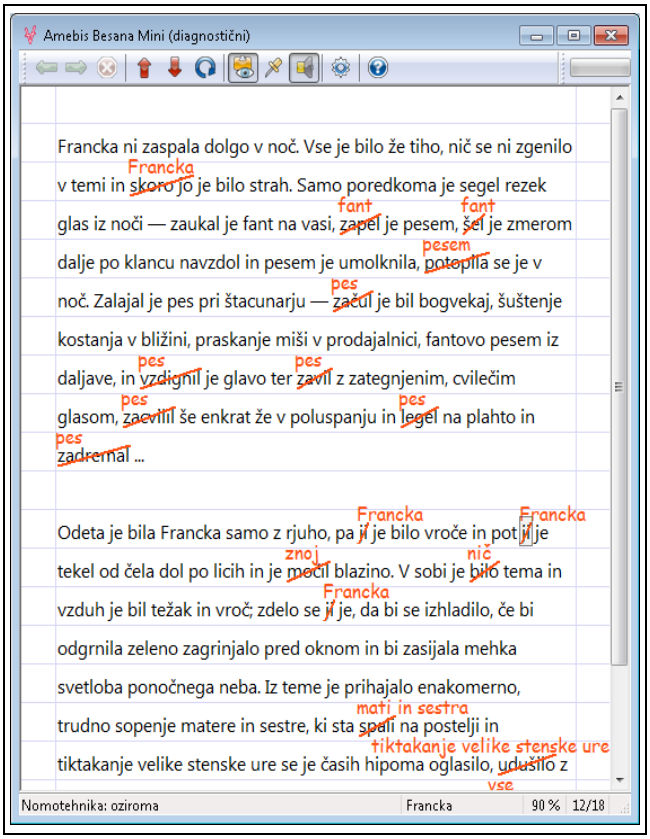
Pri katafori je zaimek pred samostalniško frazo, ki jo nadomešča. Primer za to je poved »Ker jo je zeblo, je Mojca oblekla jopico.«

Vidimo lahko, da je pri tem tipu zaimek, ki ga želimo razrešiti, v odvisniku, razrešitev sklica moramo pa poiskati v osebku glavnega stavka, ki sledi, pri čemer pa moramo paziti še na to, da je ta osebek na začetku stavka, drugače imamo težave pri primeru »Ker jo je zeblo, ji je Mojca oblekla jopico.« ali pa »Ker jo je zeblo, ju je Mojca poslala domov.«

5. Rezultati

Za lažje preizkušanje delovanja razreševanja sklicev (neposredno branje Amebisovega vmesnega jezika ni posebej preprosto, saj je bolj prilagojen temu, da ga berejo računalnik) je bila zgrajena posebna verzija slovničnega pregledovalnika Besana, in sicer z vmesnikom Besana Mini. Sklici se izpisujejo kot ena od napak, ki jih program išče (razrešitve (izpisane vedno v imenovalniku) nadomestijo osebni zaimek, če pa gre za izpust osebnega zaimka, pa nadomestijo glagol), za boljšo preglednost se izključi izpis vseh drugih napak. Na ta način se besedilo, s

katerim želimo preveriti delovanje razreševanje sklicev, le skopira v odložišče in Besana takoj izpiše najdene razrešitve, kot je primer na sliki 5.



Slika 5. Primer razreševanja sklicev na prvih dveh odstavkih romana Na klancu

Prvi rezultati (kot na primer zgornji primer) so videti obetavno, kaže pa se, da se bo treba bolj posvetiti izboljšavam analizatorja (zdaj denimo v primeru »Jetra so za vretenčarje značilen organ. Imajo osrednjo vlogo v presnovi in številne druge naloge« tako v drugi povedi ne najde izpusta osebnega zaimka, ker analizator napačno določi, da je osebek »naloge«).

Težave nastanejo tudi zato, ker sistem še ne vsebuje dovolj pomenskih omejitev, ki pomagajo pri izbiranju pravega sklica. Te omejitve bi tudi v splošnem pomagale pri razdvoumljanju, zato bo dopolnjevanje baze Ases v tej smeri zelo koristno.

Občasno se pokažejo tudi težave, ker starejši kontekst preveč vpliva na nove stavke, kar kaže, da bo smiselno preizkusiti hitrejše pozabljanje konteksta. Točne nastavitve teh parametrov pa bodo zahtevale več preizkušanja in predvsem pripravo korpusa primerov razrešenih sklicev, kar bo omogočilo hitrejše preizkušanje različic.

6. Uporaba v sistemu Piflar

Piflar je sistem za odgovarjanje na vprašanja v naravnem jeziku, ki se med drugim uporablja na Amebisovem portalu za virtualne asistente SecondEgo (http://www.secondego.com). Sistem kot osnovo uporablja Amebisov vmesni jezik, velika omejitev pa so bili zaimki v vhodnem besedilu, ker se je sistem naučil znanje z zaimki namesto z njihovimi pravimi pomeni (Holozan, 2014)

Že v (Vicedo, Ferrández, 2000) je bilo pokazano, da je razreševanje sklicev pomembno za odgovarjanje na vprašanja. Zato je bil Piflar dopolnjen s podporo za element ORI, ki je bil dodan v vmesni jezik za zapisovanje razrešenih sklicev, tako da zdaj uporablja ta element namesto originalnega jedra (JED). S to dopolnitvijo zdaj pravilno odgovarja tudi v primerih, ko je treba upoštevati sklice, kar prikazuje tabela 1, kjer so odebeljeno označena vprašanja, na katera je mogoče odgovoriti zaradi razrešenih sklicev, prej pa bi bili uporabljeni osebni zaimki oz. Piflar ni imel odgovora na vprašanje.

vprašanje	kratki odgovor	dolgi odgovor	prejšnji kratki odgovor
Ali je Miha prebral knjigo?	da	Da.	da
Kdo je prebral knjigo?	Miha	Knjigo je prebral Miha.	Miha
Kaj je Miha prebral?	knjigo	Miha je prebral knjigo.	knjigo
Ali je Miha potem pojedel kosilo?	da	Da.	/
Kdaj kosilo je pojedel Miha?	potem	Kosilo je pojedel Miha potem.	/
Kaj je Miha pojedel potem?	kosilo	Miha je pojedel kosilo potem.	/
Kdo je pojedel kosilo potem?	Miha	Kosilo je pojedel Miha potem.	on
Ali je Miha šel v Ljubljano?	da	Da.	/
Kdo je šel v Ljubljano?	Miha	V Ljubljano je šel Miha.	on
Kam je šel Miha?	v Ljubljano	Miha je šel v Ljubljano.	/
Ali je Miha naletel na Janeza?	da	Da.	/
Na koga je Miha naletel?	Janez	Miha je naletel na Janeza.	/
Kdo na Janeza je naletel?	Miha	Na Janeza je naletel Miha.	on
Ali je Miha pozdravil Janeza?	da	Da.	/
Kdo je pozdravil Janeza?	Miha	Janeza je pozdravil Miha.	on
Koga je Miha pozdravil?	Janeza	Miha je pozdravil Janeza.	/

Tabela 1: Seznam vprašanj in odgovorov, ki jih najde Piflar za primer »Miha je prebral knjigo. Potem je pojedel kosilo in šel v Ljubljano. Srečal je Janeza in ga pozdravil.«.

Podoben primer, ki pa vsebuje še prislovni zaimek, je, če imamo vhodno besedilo »Matematik Josip Plemelj se je rodil 11. decembra 1873 na Bledu. Tam je obiskoval osnovno šolo.« Na vprašanje »Kje je Josip Plemelj obiskoval osnovno šolo?« tako dobimo odgovor »Josip Plemelj je hodil v osnovno šolo na Bledu.« Težava pri teh krajevnih (in podobno časovnih) določitvah pa je, da so največkrat implicitne (se prenašajo iz prejšnjih stavkov brez izrecne uporabe prislovnih zaimkov), česar Piflar še ne zna uporabiti.

Sistem Piflar je bil dodatno dopolnjen s tem, da zna odgovarjati tudi na vprašanja, ki niso zastavljena v obliki stavka, ampak le kot posamičen stavčni člen. Tako npr. kot odziv na vprašanje »Miha« poišče dejstvo, ki vsebuje samostalniško frazo »Miha«, npr. »Miha je šel v Ljubljano.«, če ima v bazi primer iz Tabele 1. Na ta način se Piflar bolj uspešno odziva na način iskanja, na katerega so uporabniki navajeni iz običajnih iskalnikov.

Pri tovrstnih vprašanjih je možno kombinirati tudi npr. osebek in prislovno določilo. Če imamo npr. učni stavek »Isaac Newton je umrl 20. marca 1727 v Kensingtonu.«, zdaj Piflar na vprašanje »Isaac Newton 1727« odgovori: »Isaac Newton je umrl 20. marca 1727.«

Pri uporabi v sistemu Piflar pa se z bolj kompleksnimi odgovori kaže tudi to, da bo treba dopolniti tudi generator, ki prevaja vmesni jezik v naravni jezik, in sicer v smeri, da bo poskrbel za naravnejše odgovore s tem, da bo dodajal izpuste in po potrebi tudi sklice z osebnimi zaimki, da bodo odgovori zveneli bolj naravno. Zdaj npr. pri učnem besedilu »Oblaki nastajajo poleti nad večjimi ognjeniki. Tam nastanejo zato, ker se topli zrak dviga in ohlaja.« na vprašanje »Zakaj nastanejo oblaki nad večjimi ognjeniki?« odgovori »Do oblakov pride nad večjimi ognjeniki, ker se dviguje topli zrak in ker se ohlaja.«, namesto »Do oblakov nad večjimi ognjeniki pride, ker se topli zrak dviguje in ohlaja.«

7. Sklep

Razreševanje sklicev se je pokazalo kot uporabno tako v sistemu Piflar kot tudi v strojnem prevajalniku Presis (ki tako zdaj poved »Pobral sem knjigo in jo začel brati.« prevede v »I picked up a book and started to read it.« namesto v »I picked up a book and started to read her.« kot do zdaj).

Potencialna možnost uporabe je še pri iskanju po korpusih, npr. pri iskanju kolokacij, kjer bi z razširitvijo iskanja na osebne zaimke z razrešenimi sklici lahko povečali število zadetkov pri isti velikosti korpusa.

Sistemu še ne uspe razrešiti vseh sklicev, zato je še veliko možnosti za izboljšave, še posebej to velja za sklice, ki niso zaimki, niti dotaknil pa se še ni tudi bolj zapletenih povezanih sklicev (par – moški).

Razreševanje sklicev je možno izboljšati tudi z analizo diskurza, predvsem dialogov, s čimer bi se lahko bolje povezale informacije v različnih odstavkih in v premem govoru. Tak primer je npr. v sliki 6.

Rdeča kapica je vprašala volka: »Zakaj imaš tako veliko oči?« »Da te bolje vidim.«

Slika 6. Primer diskurza.

Na podlagi tega učnega besedila, bi moral biti Piflar sposoben na vprašanje »Zakaj ima volk tako velike oči?« odgovoriti z »Volk ima tako velike oči, da bolje vidi Rdečo kapico.«

Pogoj za nadaljnji razvoj pa bo verjetno tudi priprava korpusa primerov razrešenih sklicev, ki bi omogočil hitro primerjavo delovanja različnih postopkov, pri pripravi takega korpusa pa bi bilo pomembno, da se ne omeji le na osebne zaimke, ampak se označijo tudi druge vrste sklicev, da bo tak korpus uporaben tudi za preizkušanje razreševanja bolj zapletenih vrst sklicev.

8. Literatura

Brennan, S. E., Friedman M. W., Pollard C. J., 1987. A centering approach to pronouns. V *Proceedings of the 25th Annual Meeting of the ACL*. Stanford. 155–162.

Erjavec, T., Krek, S., 2008: Oblikoskladenjske specifikacije in označeni korpusi JOS, V T. Erjavec, J. Žganec Gros (ur.), *Zbornik 6. konference Jezikovne tehnologije 2008*. Ljubljana: IJS.

Ge, N., Hale, J., Charniak E., 1998. A statistical approach to anaphora resolution. V *Proceedings of the Sixth Workshop on Very Large Corpora*.

Hobbs, J. R. 1978. Resolving pronoun references. V Barbara J. Grosz, Karen Spärck-Jones, and Bonnie Lynn Webber (ur.), *Readings in Natural Language Processing*. Morgan Kaufmann Publishers, Los Altos. 339–352.

Holozan, P., 2011. *Samodejno izdelovanje besedilnih logičnih nalog v slovenščini.* Magistrsko delo, Univerza v Ljubljani, Fakulteta za računalništvo in informatiko.

Holozan, P., 2014. Piflar: sistem za učenje in odgovarjanje na vprašanja v naravnem jeziku. V M. Orel in S. Jurjevčič (ur.), *MEDNARODNA konferenca InfoKomTeh.* Polhov Gradec: Eduvision.

Joshi, A. K., Kuhn, S., 1979. Centered logic: The role of entity centered sentence representation in natural language inferencing. V *Proceedings of the International Joint Conference on Artificial Intelligence*, Tokyo. 435–439.

Lappin, S., Leass. H. J., 1994. An algorithm for pronominal anaphora resolution. *Computatinal Linguistics*, 20(4). 535–561.

McShane, M. Beale, S., Nirenburg, S., 2010. Reference Resolution Supporting Lexical Disambiguation. V zborniku *2010 IEEE Fourth International Conference on Semantic Computing*. Los Alamitos: IEEE Computer Society.

Mitkov R., 1999. *Anaphora Resolution: The State Of The Art.* Working paper, University of Wolverhampton.

Mitkov, R., Evans C., Orasan, C., 2002. A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. V *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, Mexico City, Mexico, February, 17 – 23.

Němčík, V., 2006. *Anaphora Resolution.* Magistrsko delo, Masarykova universita, Fakulta informatiky.

Vicedo, J. L., Ferrández, A., 2000. Importance of Pronominal Anaphora resolution in Question Answering systems. V *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL}*. 555–562.

Alp-ULj Speaker Recognition System for the NIST 2014 i-Vector Challenge

Boštjan Vesnicer,* Jerneja Žganec-Gros,* Simon Dobrišek† and Vitomir Štruc†

*Alpineon d.o.o.
Ulica Iga Grudna 15, SI-1000 Ljubljana
{bostjan.vesnicer,jerneja.gros}@alpineon.si

†Faculty of Electrical Engineering
University of Ljubljana
Tržaška cesta 25, SI-1000 Ljubljana
{simon.dobrisek,vitomir.struc}@fe.uni-lj.si

Abstract

I-vectors enable a fixed-size compact representation of speech signals of arbitrary durations. In recent years they have become the state-of-the-art representation of speech signals in text-independent speaker recognition. For practical reasons most systems assume that the i-vector estimates are highly reliable. However, this assumption is valid only in the case when i-vectors are extracted from recordings of sufficient length, but for short recordings the assumption does not hold any more. To address the problem of duration variability we propose a simple duration-based preprocessing weighting scheme that accounts for different reliability of i-vector estimates. We evaluate the proposed approach in the scope of NIST 2014 i-vector machine learning challenge, where we achieved competitive results.

Sistem za prepoznavo govorcev Alp-ULj s prireditve "NIST 2014 i-Vector Challenge"

I-vektorji omogočajo zgoščeno predstavitev govornih signalov poljubne dolžine v obliki vektorjev fiksne razsežnosti. V zadnjih letih so postali ena izmed najuspešnejših tehnologij na področju prepoznave govorcev. Zaradi praktičnih razlogov ponavadi predpostavimo, da je ocena i-vektorjev zelo zanesljiva. Ta predpostavka velja le v primeru, ko i-vektor ocenimo iz dovolj dolgega govornega posnetka, medtem ko je pri posnetkih krajše dolžine ta predpostavka v veliki meri kršena. V prispevku predlagamo posebno metodo predobdelave, v kateri na enostaven način upoštevamo dolžino posnetkov, iz katerih smo i-vektorje ocenili. Predlagano rešitev smo ovrednotili v okviru prireditve "NIST 2014 i-Vector Challenge", na kateri smo dosegli vzpodbudne rezultate.

1. Introduction

The area of speaker recognition has made significant progress over recent years. Today, recognition systems relying on so-called i-vectors, introduced in (Dehak et al., 2011), have emerged as the de-facto standard in this area. Most of the existing literature on i-vector-based speaker recognition focuses on recognition problems, where the i-vectors are extracted from speech recordings of sufficient length. The length of the recordings is predefined by the speech corpus used for the experimentation and typically does not drop below a length that would cause problems to the recognition techniques. In practical applications, however, speaker recognition systems often deal with i-vectors extracted from short recordings, which may be estimated less reliably than i-vectors extracted from recordings of sufficient length.

The problem of duration variability is known to be one of importance for practical speaker-recognition applications and has also been addressed to a certain extent in the literature in the context of i-vector-based speaker-recognition systems, e.g. (Sarkar et al., 2012; Kanagasundaram et al., 2011; Hasan et al., 2013a; Mandasari et al., 2011; Garcia-Romero and McCree, 2013; Kenny et al., 2013; Cumani et al., 2013; Kanagasundaram et al., 2014; Hasan et al., 2013b; Stafylakis et al., 2013). The most recent solutions of the duration-variability problem, e.g. (Garcia-Romero and McCree, 2013; Kenny et al., 2013; Cumani et al., 2013) do not treat i-vectors as point estimates of the hidden variables in the eigenvoice model, but rather as random vectors. In this slightly different perspective, the i-vectors appears as posterior distributions, parameterized by the posterior mean and the posterior covariance matrix. Here, the covariance matrix can be interpreted as a measure of the uncertainty of the point estimate that relates to the duration of the speech recording used to compute the i-vectors.

In this paper we propose a slightly different approach and try to compensate for the problem of duration variability of the speech recordings through weighted statistics. Typically, feature-transformation techniques commonly used in the area of speaker recognition, such as principal component analysis (PCA) or within-class covariance normalization (WCCN) estimate the covariance matrices and sample means by considering the contribution of each available i-vector equally in the statistics, regardless of the fact that the i-vectors may be estimated unreliably. To address this point, we associate with every i-vector a weight that is proportional to the duration of the speech recording from which the i-vector was extracted. This weight is then

This work was supported in parts by the national research program P2-0250(C) Metrology and Biometric Systems, the European Union's Seventh Framework Programme (FP7-SEC-2011.20.6) under grant agreement number 285582 (RESPECT), the Eureka project S-Verify (contract No. 2130-13-090145) and by the European Union, European Regional Fund, within the scope of the framework of the Operational Programme for Strengthening Regional Development Potentials for the Period 2007-2013, contract No. 3330-13-500310 (eCall4All). The authors additionally appreciate the support of COST Actions IC1106 and IC1206.

used to control the impact of a given i-vector to the overall statistics being computed. The described procedure can be applied to any feature transformation technique and results in duration-weighted techniques that should lead to better estimates of the feature transforms.

We evaluate the proposed weighting scheme in the scope of the NIST 2014 i-vector machine learning challenge (IVC). The goal of the challenge is to advance the state-of-technology in the area of speaker recognition by providing a standard experimental protocol and pre-computed i-vectors for experimentation. Based on the data provided by the challenge, we show that it is possible to apply the proposed weighting scheme to supervised as well as unsupervised feature-transformation techniques and that in both cases performance gains can be expected. With our best performing (duration-weighted) system we managed to achieve a minimal decision-cost-function (DCF) value of 0.280, a 27% relative improvement over the baseline system.

2. Prior work

Two of the most frequently used classification methods in i-vector-based speaker recognition are the cosine similarity (Dehak et al., 2010) and probabilistic linear discriminant analysis (PLDA), independently developed for face (Prince and Elder, 2007; Li et al., 2012) and speaker recognition (Kenny, 2010). Since its introduction, the PLDA model has been extended in different ways, e.g. the underlying Gaussian assumption have been relaxed (Kenny, 2010), the parameters of the model have been treated as random variables (Villalba and Brummer, 2011) and an extension to the mixture case has been proposed as well (Senoussaoui et al., 2011).

Before given to the classifier, i-vectors are usually preprocessed in various ways. Common preprocessing methods include whitening (PCA), linear discriminant analysis (LDA) and within-class covariance normalization (WCCN), which can be applied in combination. Another important preprocessing step is length normalization, as it turns out (Garcia-Romero and Espy-Wilson, 2011) that length normalization brings the i-vectors closer to a normal distribution and therefore provides for a better fit with the assumptions underlying Gaussian PLDA.

3. Duration-based weighting

In this section we introduce our duration-dependent weighting scheme. We assume that the front-end processing of the speech recording has already been conducted and that all we have at our disposal is a set of extracted i-vectors and a single item of metadata in the form of the duration of the recording from which a given i-vector was extracted (NIST, 2014). Under the presented assumptions the solutions to the problem of duration variability that treat the i-vectors as random variables characterized by a posterior distribution, such as those presented in (Garcia-Romero and McCree, 2013; Kenny et al., 2013; Cumani et al., 2013), are not applicable.

The basic step in computing the feature transform for most feature-extraction (or feature-transformation) techniques (e.g., PCA, WCCN, NAP, etc.) is the calculation of the sample mean and scatter (or covariance) matrix. Given some training i-vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, with $\mathbf{x}_i \in \mathbb{R}^m$ and $i = 1, 2, \ldots, n$, the sample mean $\boldsymbol{m}$ and scatter matrix $\boldsymbol{S}$ can be calculated by the following formulas:

$$\boldsymbol{m} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \tag{1}$$

and

$$\boldsymbol{S} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{m})(\mathbf{x}_i - \boldsymbol{m})^{\mathrm{T}}. \tag{2}$$

The definition of the sample mean and scatter matrix in Eqs. (1) and (2) assume that all the training vectors $\mathbf{x}_i$ ($i = 1, 2, \ldots, n$) are equally reliable and are, therefore, given equal weights when computing the mean and covariance matrix. While such an interpretation of the equations is (most likely) valid if the training vectors are computed from speech recordings of sufficient length, this may not be true if some of the vectors are extracted from short recordings. In this case, some of the training vectors are unreliable and should not contribute equally to the computed statistics.

To account for the above observation we propose to multiply the contribution of each i-vector in Eqs. (1) and (2) by the weight which corresponds to the duration of the recording from which the vector was extracted. This modification gives the following formulas for the weighted mean $\boldsymbol{m}_{\boldsymbol{w}}$ and weighted scatter matrix $\boldsymbol{S}_{\boldsymbol{w}}$:

$$\boldsymbol{m}_{\boldsymbol{w}} = \frac{1}{T} \sum_{i=1}^{n} t_i \mathbf{x}_i \tag{3}$$

and

$$\boldsymbol{S}_{\boldsymbol{w}} = \frac{1}{T} \sum_{i=1}^{n} t_i (\mathbf{x}_i - \boldsymbol{m}_{\boldsymbol{w}})(\mathbf{x}_i - \boldsymbol{m}_{\boldsymbol{w}})^{\mathrm{T}}, \tag{4}$$

where $T = \sum_{i=1}^{n} t_i$.

Note that the presented weighting scheme reduces to the (non-weighted) standard version if the speech recordings, from which the training vectors are extracted, are of the same length. If this is not the case, the presented weighting scheme gives larger emphasis to more reliably estimated i-vectors. In the remainder, we present modifications of two popular feature-transformation techniques based on the presented weighting scheme, namely, PCA and WCCN. We first briefly describe the theoretical basis of both techniques and then show, how they can be modified based on the presented statistics.

3.1. Principal component analysis

Principal component analysis (PCA) is a powerful statistical learning technique with applications in many different areas, including speaker verification. PCA learns a subspace from some training data in such a way that the learned basis vectors correspond to the maximum variance directions present in the original training data (V. Štruc and Pavešić, 2008). Once the subspace is learned, any given feature vector can be projected into the subspace to be processed further or to be used with the selected scoring procedure. In state-of-the-art speaker-verification systems the feature vectors used with PCA typically take the form of

i-vectors, which after processing with the presented technique are fed to a scoring technique, based on which identity inference is conducted.

Formally PCA can be defined as follows. Given a data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n], \mathbf{x}_i \in \mathbb{R}^m$ containing in its columns n training vectors $\mathbf{x}_i$, for $i = 1, 2, \ldots, n$, PCA computes a subspace basis $\mathbf{U} \in \mathbb{R}^{m \times d}$ by factorizing of the covariance matrix $\mathbf{\Sigma}$ of the vectors in $\mathbf{X}$ into the following form:

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T, \tag{5}$$

where $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d], \mathbf{u}_i \in \mathbb{R}^m$ denotes an orthogonal eigenvector vector matrix (i.e., the projection basis) and $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_d, \}$ stands for a diagonal eigenvalue matrix with the eigenvalues arranged in decreasing order. Note that if $\mathbf{\Sigma}$ is full-rank the maximum possible value for the subspace dimensionality is $d = n$, if the covariance matrix is not full-rank the upper bound for d is defined by the number of non-zero eigenvalues in $\mathbf{\Lambda}$. In practice, the dimensionality of the PCA subspace d is an open parameter and can be selected arbitrarily (up to the upper bound).

Based on the computed subspace basis, a given feature vector $\mathbf{x}$ can be projected onto the $d-$dimensional PCA subspace using the following mapping:

$$\mathbf{y} = \mathbf{U}^T(\mathbf{x} - \boldsymbol{\mu}), \tag{6}$$

where $\mathbf{y} \in \mathbb{R}^d$ stands for the PCA transformed feature vector.

Commonly, the above transformation is implemented in a slightly different form, which next to projecting the given feature vector $\mathbf{x}$ into the PCA subspace, also whitens the data:

$$\mathbf{y} = (\mathbf{U}\mathbf{\Lambda}^{-1/2})^T(\mathbf{x} - \boldsymbol{\mu}). \tag{7}$$

3.2. Within-class covariance normalization

Within-Class Covariance Normalization (WCCN) is a feature transformation technique originally introduced in the context of Support Vector Machine (SVM) classification (Hatch and Stolcke, 2006). WCCN can under certain conditions be shown to minimize the expected classification error by applying a feature transformation on the data that as a result whitens the within-class scatter matrix of the training vectors. Thus, unlike PCA, WCCN represents a supervised feature extraction/transformation technique and requires the training data to be labeled. In state-of-the-art speaker verification systems, the feature vectors used with WCCN typically represent i-vectors (or PCA-processed i-vectors) that after the WCCN feature transformation are subjected to a scoring procedure.

Typically WCCN is implemented as follows. Consider a data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n], \mathbf{x}_i \in \mathbb{R}^m$ containing in its columns n training vectors $\mathbf{x}_i$, for $i = 1, 2, \ldots, n$, and let us further assume that these vectors belong to N distinct classes $C_1, C_2, \ldots, C_N$ with the j-th class containing n_j samples and $n = \sum_{j=1}^{N} n_j$. WCCN computes the transformation matrix based on the following Cholesky factorization:

$$\mathbf{\Sigma}_w^{-1} = \mathbf{L}\mathbf{L}^{\mathrm{T}}, \tag{8}$$

where $\mathbf{L}$ and $\mathbf{L}^T$ stand for the lower and upper triangular matrices, respectively, and $\mathbf{\Sigma}_w^{-1}$ denotes the inverse of the within-class scatter matrix computed from the training data.

Once computed, the WCCN transformation matrix $\mathbf{L}$ can be used to transform any given feature vector $\mathbf{x}$ based on the following mapping:

$$\mathbf{y} = \mathbf{L}^{\mathrm{T}}\mathbf{x}, \tag{9}$$

where $\mathbf{y} \in \mathbb{R}^m$ stands for the transformed feature vector.

The weighted version of the WCCN transform can be obtained by replacing the standard withing-class scatter matrix with the weighted one.

4. The I-vector challenge

We evaluate the feasibility of the proposed duration-weighted scheme in the scope of IVC. In this section we provide some basic information on the challenge, present the experimental protocol and define the performance metric used to assess the recognition techniques.

4.1. Challenge description

The single task of IVC is that of speaker detection, i.e., to determine whether a specified speaker (the target speaker) is speaking during a given segment of conversational speech. The IVC data is given in the form of 600-dimensional i-vectors, divided into disjoint development and evaluation sets. The development set consists of 36,572 (unlabeled) i-vectors, while the evaluation set consists of 6,530 target i-vectors belonging to 1,306 target speakers (5 i-vectors per speaker) and 9,643 test i-vectors of a unknown number of speakers. Note that no explicit information is provided on whether the 1,306 speakers are distinct or not. Hence, it is possible that some of the target identities are duplicated.

The experimental protocol of IVC defines that a total of 12,582,004 experimental trials need to be conducted, where each trial consists of matching a single i-vector from the 9,643 test vectors against a given target model constructed based on the five target i-vectors belonging to the targeted speaker. It should be noted that — according to the rules (NIST, 2014) — the output produced for each trial must be based (in addition to the development data) solely on the training and test segment i-vectors provided for that particular trial, while the i-vectors provided for other trials may not be used in any way.

The durations of the speech segments used to compute the i-vectors for IVC are sampled from a log-normal distribution with a mean of 39.58 seconds. This suggests that methods that take the uncertainty of the i-vectors due to duration variability into account should be effective in the challenge. However, since the only information provided with each i-vector is the duration of the speech recording used to compute the corresponding i-vector, techniques exploiting the posterior covariance, such as (Garcia-Romero and McCree, 2013; Kenny et al., 2013; Cumani et al., 2013), are not feasible. Nevertheless, we expect that performance improvements should be possible by augmenting the information contained in the i-vectors with duration information in one way or another.

5. Experiments and results

5.1. Experimental setup

The experiments presented in the remainder are conducted in accordance with the experimental protocol defined for the i-vector challenge and presented in Section 4.1.. The processing is done on a personal desktop computer using Matlab R2010b and the following open source toolboxes:

- the PhD toolbox (Štruc and Pavešić, 2010; Štruc, 2012)[1], which among others features implementations of popular dimensionality-reduction techniques;
- the Bosaris toolkit (Brummer and de Villiers, 2011)[2], which contains implementations of score calibration, fusion and classification techniques;
- the Liblinear library (with the Matlab interface) (Fan et al., 2008)[3], which contains fast routines for training and deploying linear classifiers such as linear SVMs or logistic-regression classifiers.

All the experiments presented in the next sections can easily be reproduced using the above tools and functions.

5.2. Experiments with PCA

Our duration-dependent weighting scheme is based on the assumption that not all the available i-vectors are computed from speech recordings of the same length and are, therefore, not equally reliable. If the i-vectors are computed from recordings of comparable length, the weighting scheme would have only little effect on the given technique, as similar weights would be assigned to all the statistics and the impact of the weighting would basically be lost. On the other hand, if the i-vectors are computed from speech recordings of very different lengths, our weighting scheme is expected to provide more reliable results, as more reliable i-vectors are given larger weights when computing statistics for the given speaker-verification technique.

To assess our weighting scheme we first implement the baseline technique defined for the i-vector challenge and use the baseline performance for comparative purposes. Note that IVC defines a PCA-based system used together with cosine scoring as its baseline. Specifically, the baseline system consists of the following steps (NIST, 2014)

- estimation of the global mean and covariance based on the development data,
- centering and whitening of all i-vectors based on PCA (see Eq. 7),
- projecting all i-vectors onto the unit sphere (i.e., length normalization: $\mathbf{x} \leftarrow \frac{\mathbf{x}}{\sqrt{\mathbf{x}^T\mathbf{x}}}$),
- computing models by averaging the five target i-vectors of each speaker and normalizing the result to unit L_2 norm, and
- scoring by computing inner products between all models and test i-vectors.

[1] http://luks.fe.uni-lj.si/sl/osebje/vitomir/face_tools/PhDface
[2] https://sites.google.com/site/bosaristoolkit
[3] http://www.csie.ntu.edu.tw/~cjlin/liblinear

Table 1: *Effect of the proposed weighting scheme on the baseline system defined for IVC. The Table shows minDCF values achieved by the baseline and weighted baseline systems as returned by the web-platform of the IVC as well as the relative change (in%) in the minDCF value, achieved with the weighting.*

Technique	Baseline	Weighted baseline	minDCF_{rel}
Score	0.386	0.372	3.63%

Table 2: *Effect of excluding samples from the development set of the IVC data on the performance of the baseline and weighted baseline systems. The exclusion criterion is a threshold on the duration of the recording used to compute the i-vectors. The Table shows minDCF values as returned by the web-platform of the IVC.*

Exclusion criterion	$< 10s$	$< 15s$	$< 20s$	$< 25s$
Baseline	0.385	0.381	0.379	0.377
Weighted	0.372	0.371	0.371	0.371

In our first series of experiments, we modify the baseline system by replacing the PCA step (second bullet) with our duration-weighted version of the PCA. We provide the comparative results in terms of the minDCF values in Table 1. Here, the last column denotes the relative change in the minDCF value measured against the baseline:

$$\text{minDCF}_{rel} = \frac{\text{minDCF}_{base} - \text{minDCF}_{test}}{\text{minDCF}_{base}}, \qquad (10)$$

where minDCF_{base} stands for the minDCF value of the baseline system and minDCF_{test} stands for the minDCF value achieved by the currently assessed system.

Note that the proposed weighting scheme results in a relative improvement of 3.63% in the minDCF value over the baseline. This result suggests that a performance improvement is possible with the proposed weighting scheme, but a more detailed analysis of this results is still of interest. For this reason we examine the behavior of the baseline and weighted baseline techniques with respect to a smaller development set, where i-vectors computed from shorter recordings are excluded from the estimation of the global mean and covariance. Based on this strategy, we construct four distinct development sets with the first excluding all the i-vectors with the associated duration shorter than 10s, the second excluding all the i-vectors with the associated duration shorter than 15s, the third excluding all the i-vectors with the associated duration shorter than 20s, and the last excluding all i-vectors with the associated duration shorter than 25s. The baseline and weighted baseline technique are then trained on the described development sets. The results of this series of experiments are presented in Table 2.

Note that by excluding vectors from the development set, the baseline technique gradually improves in perfor-

mance as more and more of the unreliable i-vectors are excluded from training. Continuing this procedure would clearly turn the trend around and the minDCF values would start getting worse, as too much information would be discarded. The weighted baseline system, on the other hand, ensures minDCF values comparable to those that were achieved when the entire development set was used for the training. This result again suggests that duration variability is addressed quite reasonably with the proposed weighting scheme.

5.3. Experiments with WCCN

In the next series of experiments we assess the performance of WCCN-based recognition systems. As a baseline WCCN system, we implement a similar processing pipeline as presented for the IVC baseline technique in the previous section, but add an additional step, which after whitening with PCA also whitens the within-class covariance matrix using WCCN. All the remaining steps of our WCCN-based baseline stay the same including length normalization, model construction and scoring. Whenever using the weighted version of WCCN we also use the weighted version of PCA in the experiments.

To further improve upon the baseline, we implement a second group of WCCN-based systems, where the cosine-based scoring procedure is replaced with a logistic-regression classifier and the length normalization is removed from the processing pipeline. With this approach all five target i-vectors of a given speaker are considered as positive examples of one class, while 5,000 i-vectors most similar to the given target speaker are considered as negative examples of the second class. Based on this setup a binary classifier is trained for each target speaker, resulting in a total of 1,306 classifiers for the entire IVC data.

Before we turn our attention to the experimental results, it has to be noted that unlike PCA, which is an unsupervised technique, WCCN represents a supervised feature transformation techniques, which requires that all i-vectors comprising the development data are labeled. Unfortunately, the development data provided for the i-vector challenge is not labeled nor is the number of speakers present in the data known. To be able to apply supervised algorithms successfully we need to generate labels in an unsupervised manner by applying an appropriate clustering algorithm (Senoussaoui et al., 2014). Clustering will, however, never be perfect in practice, so the errors (utterances originated from the same speaker can be attributed to different clusters or utterances from different speakers can be attributed to the same cluster) are inevitable. Although there exists some evidence that labeling errors can degrade the recognition performance (seen as a bending of the DET curve), it is not completely obvious how sensitive different methods are with respect to those errors.

Since the selection of an appropriate clustering technique is (clearly) crucial for the performance of the supervised feature transformation techniques, we first run a series of preliminary experiments with respect to clustering and elaborate on our main findings. The basis for our experiments is whitened i-vectors processed with the (PCA-based) baseline IVC system. We experiment with different clustering techniques (i.e., k-means, hierarchical clustering, spectral clustering, mean-shift clustering, k-medoids and others), using different numbers of clusters and different (dis-)similarity measures (i.e., Euclidian distances and cosine similarity measures). The results of our preliminary experiments suggest the cosine similarity measure results in i-vector labels that ensure better verification performance than the labels generated by the Euclidian distance (with the same number of clusters). Despite the fact that several alternatives have been assessed, classical k-means clustering ensures the best results in our experiments and was, therefore, chosen as the clustering algorithm for all of our main experiments. Based on our preliminary experiments, we select the k-means clustering algorithm with the cosine similarity measure for our experiments with WCCN and run it on the development data. We set the number of clusters to 4,000, which also ensured the best results during our preliminary experimentation.

Table 3: *Effect of the proposed weighting scheme on our WCCN-baseline system. The Table shows minDCF values achieved by the baseline and weighted baseline WCCN systems as returned by the web-platform of the IVC as well as the relative change (in%) in the minDCF value, achieved with the weighting.*

Technique	Baseline	Weighted	minDCF$_{rel}$
Cosine	0.461	0.447	3.04%
Logistic	0.304	0.294	3.29%

The results of the WCCN-based series of experiments are presented in Table 3. Here, the relative change in the minDCF value is measured against the WCCN baseline. The first thing to notice is that with cosine scoring the WCCN-baseline systems (weighted and non-weighted) result in significantly worse minDCF values. However, when the scoring procedure is replaced with a logistic-regression classifier, this changes dramatically. In this situation, the WCCN-based system becomes highly competitive and in the case of the weighted system result in a minDCF value of 0.294. All in all, the weighting scheme seems to ensure a consistent improvement over the non-weighted case of around 3%. For the sake of completeness we need to emphasize that the best score we managed to achieve with a PCA-based system, when using a logistic-regression classifier was 0.326.

5.4. Comparative assessment

For the i-vector challenge we further tuned our best performing recognition system (i.e., the weighted version of our WCCN-system) to achieve even lower minDCF values. After implementing several additional steps we managed to reduce the minDCF value of our system to 0.280 by the time of writing. Specifically, the following improvements were implemented:

- duration was added as an additional feature to the i-vectors to construct 601 dimensional vectors before any processing,

- the clustering was improved by excluding clusters with a small fisher-score,
- the entire development set was used as negative examples when training the classifiers, and
- a second set of classifiers was trained on the test vectors and then used to classify the target vectors; the mean score over a given target speaker was then combined with the score computed based on the classifier trained on the target identity.

6. Conclusions

We have presented a duration-based weighting scheme for feature transformation techniques used commonly in an i-vector based speaker-recognition system. We have applied the scheme on two established transformation techniques, namely, principal component analysis and within-class covariance normalization. We have assessed the duration-weighted techniques in the scope of the NIST i-vectormachine learning challenge and achieved very competitive results. As part of our future work, we plan to evaluate the possibility of using a similar scheme with probabilistic linear discriminant analysis as well.

7. References

N. Brummer and E. de Villiers. 2011. The BOSARIS toolkit user guide: Theory, algorithms and code for surviving hte new dcf. In *NIST SRE'11 Analysis Workshop*, Atlanta, USA, December.

S. Cumani, O. Plchot, and P. Laface. 2013. Probabilistic linear discriminant analysis of i-vector posterior distributions. In *Proc. ICASSP*, Vancouver, Canada.

N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny. 2010. Cosine similarity scoring without score normalization techniques. In *Proc. Odyssey*, Brno.

N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. 2011. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 19(4):788–798.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

D. Garcia-Romero and C. Y. Espy-Wilson. 2011. Analysis of i-vector length normalization in speaker recognition systems. In *Proceedings of Interspeech*, Florence, Italy.

D. Garcia-Romero and A. McCree. 2013. Subspace-constrained supervector PLDA for speaker verification. In *Proc. Interspeech*, Lyon, France.

T. Hasan, S.O. Sadjadi, G. Liu, N. Shokouhi, H. Boril, and J.H. Hansen. 2013a. Crss systems for 2012 nist speaker recognition evaluation. In *Proc. ICASSP*, Vancoucer, Canada.

T. Hasan, R. Saeidi, J. H. L. Hansen, and D. A. van Leeuwen. 2013b. Duration mismatch compensation for i-vector based speaker recognition systems. In *Proc. ICASSP*.

A. Hatch and A. Stolcke. 2006. Generalized linear kernels for one-versus-all classification: application to speaker recognition. In *Proc. ICASSP*, Toulouse, France.

A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason. 2011. I-vector based speaker recognition on short utterances. In *Proc. Interspeech*, pages 2341–2344, Florence, Italy.

A. Kanagasundaram, D. Dean, S. Sridharan, J. Gonzalez-Dominguez, J. Gonzalez-Rodriguez, and D. Ramos. 2014. Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques. *Speech Communication*, 59(April):69–82.

P. Kenny, T. Stafylakis, P. Ouellet, J. Alam, and P. Dumouchel. 2013. PLDA for speaker verification with utterances of arbitrary duration. In *Proc. ICASSP*, Vancouver, Canada.

P. Kenny. 2010. Bayesian speaker verification with heavy-tailed priors. In *Proc. Odyssey*, Brno, Czech Republic.

P. Li, Y. Fu, U. Mohammed, J.H. Elder, and S. J.D. Prince. 2012. Probablistic models for inference about identity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):144–157.

M.I. Mandasari, M. McLaren, and D.A. van Leeuwen. 2011. Evaluation of i-vector speaker recognition systems for forensics application. In *Proc. Interspeech*, pages 21–24, Florance, Italy.

NIST. 2014. The 2013-2014 speaker recognition i-vector machine learning challenge. Available online.

S. J. D. Prince and J. H. Elder. 2007. Probabilistic linear discriminant analysis for inferences about identity. In *Proc. ICCV*, Rio de Janeiro, Brazil.

A. Sarkar, D. Matrouf, P. Bousquet, and J. Bonastre. 2012. Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification. In *Proc. Interspeech*, Portland, OR, USA.

M. Senoussaoui, P. Kenny, N. Brummer, and P. Dumouchel. 2011. Mixture of PLDA models in i-vector space for gender independent speaker recognition. In *Proc. Interspeech*, Florence, Italy.

M. Senoussaoui, P. Kenny, T. Stafylakis, and P. Dumouchel. 2014. A study of the cosine distance-based mean shift for telephone speech diarization. *IEEE Transaction on Audio, Speech and Language Processing*, 22(1).

T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and P Dumouchel. 2013. Text-dependent speaker recognition using plda with uncertainty propagation. In *Proc. Interspeech*.

F. Mihelič V. Štruc and N. Pavešić. 2008. Combining experts for improved face verification performance. In *Proceedings of the International Electrotechnical and Computer Science Conference (ERK)*, pages 233–236, Portorož, Slovenia.

J. Villalba and N. Brummer. 2011. Towards fully bayesian speaker recognition: Integrating out the between speaker covariance. In *Proc. Interspeech*, Florence, Italy.

V. Štruc and N. Pavešić. 2010. The complete Gabor-Fisher classifier for robust face recognition. *EURASIP Advances in Signal Processing*, 2010:26.

V. Štruc. 2012. The PhD face recognition toolbox: toolbox description and user manual. Available online.

Razpoznavalnik tekočega govora UMB Broadcast News 2014: kakšno vlogo igra velikost učnih virov?

Andrej Žgank, Gregor Donaj, Mirjam Sepesy Maučec

Inštitut za elektroniko in telekomunikacije, Fakulteta za elektrotehniko, računalništvo in informatiko,
Univerza v Mariboru
Smetanova ul. 17, 2000 Maribor
andrej.zgank@um.si, gregor.donaj@um.si, mirjam.sepesy@um.si

Povzetek

V članku bomo predstavili rezultate analize, kako na uspešnost razpoznavanja tekočega govora vpliva velikost učnih virov, ki jih uporabimo pri izdelavi akustičnih in jezikovnih modelov. Analizo smo zasnovali na sistemu za razpoznavanje slovenskega tekočega govora UMB Broadcast News, katerega domena so dnevno-informativne televizijske oddaje. Za izhodišče smo uporabili verzijo sistema predstavljeno leta 2010, ki smo jo nadgradili z akustičnimi in jezikovnimi modeli naučenimi na dodatnih virih. Z nadgrajenim sistemom smo dosegli najboljšo pravilnost razpoznavanja besed 73,30%, kar sicer predstavlja majhno izboljšanje, vendar je bilo za to potrebno uporabiti bistveno obsežnejše vire. Z večanjem obsega virov se izboljšanje sistema zmanjšuje. Zato je razen obsega uporabljenih virov smiselno razmišljati tudi o učinkovitejših, jeziku prilagojenih postopkih razpoznavanja govora.

UMB Broadcast News 2014 continuous speech recognition system: what is the influence of language resources' size?

This paper presents the results of an analysis, how the size of language resources for training acoustic and language models influences the speech recognition accuracy. The Slovenian continuous speech recognition system UMB Broadcast News was used for the experiments. Its speech recognition domain are TV news shows. As baseline, the system presented in 2010 was used. The acoustic and language models training procedure for the experiments applied additional language resources. The improved speech recognition system achieved 73.30% word accuracy. The best speech recognition result presents a small accuracy improvement but significantly larger language resources were needed to achieve this. Enlarging language resources after a certain size brings only small improvements. This indicates that further research into language adapted methods for speech recognition is needed.

1. Uvod

Za razliko od nekaterih drugih sodobnih informacijsko komunikacijskih tehnologij, ima pri govornih tehnologijah še vedno zelo pomembno vlogo jezik. Slovenščina sodi v kategorijo visoko pregibnih jezikov, hkrati pa je zaradi relativno majhnega števila govorcev tudi tržno nezanimiva. Tako jo lahko z vidika trenutnega razvoja področja razpoznavanja govora ponovno prištevamo med jezike s premajhnim obsegom govornih virov, ki so na voljo za uporabo, hkrati pa doseženi rezultati razpoznavanja tekočega govora zaostajajo za rezultati doseženimi za nekatere jezike z velikim številom govorcev.

Za doseganje boljših rezultatov razpoznavanja govora se običajno uporabljata dve metodi. Prva je povečevanje obsega virov uporabljenih za učenje akustičnih in jezikovnih modelov. Druga metoda pa je uporaba izboljšanih algoritmov, ki lahko dodatno upoštevajo katero izmed specifik jezika (Rotovnik et al., 2007; Dobrišek & Mihelič, 2010), kot sta na primer v slovenskem jeziku visoka pregibnost besed in relativno prosti vrstni red besed v stavku.

Raziskovalno delo na področju razpoznavanja tekočega slovenskega govora je poskušalo slediti obema možnima metodama za izboljšanje rezultatov. Konec aprila 2014 je v Ljubljani potekala delavnica, kjer so predstavili rezultate evropskega projekta za avtomatsko podnaslavljanje in prevajanje predavanj, kjer je bila kot eden izmed jezikov vključena tudi slovenščina. V razpravi po predstavitvi je bila ena izmed tem tudi, ali bi za doseganje boljših rezultatov za slovenski jezik zadostovalo samo povečevanje učnih virov ali pa so potrebni tudi kakšni posebni algoritmi, ki naslavljajo značilnosti visoko pregibnih jezikov.

V članku[1] bomo poskusili analizirati, koliko lahko k izboljšanju rezultatov razpoznavanja govora doprinese povečevanje učnih virov, v našem primeru govorne baze za učenje akustičnih modelov in besedilnega korpusa za učenje statističnega jezikovnega modela. Eksperimente smo zasnovali na sistemu avtomatskega razpoznavanja tekočega slovenskega govora UMB Broadcast News (Žgank & Sepesy Maučec, 2010; Žgank et al., 2008), ki razpoznava govor v dnevno-informativnih televizijskih oddajah. Povečanje učnih virov bomo izvedli izključno z uporabo dodatnih slovenskih virov. V preteklosti smo že izvajali križno-jezične eksperimente (Žgank et al., 2004/2) - sicer samo z razpoznavanjem izoliranih in vezanih besed - kjer smo za razpoznavanje slovenskega govora uporabili akustične modele naučene na govorni bazi slovaškega jezika, ki predstavlja enega izmed slovenščini sorodnejših jezikov. Rezultati takšnega križno-jezičnega eksperimenta so pokazali, da je za preproste scenarije sicer možno doseči zelo dobre rezultate, z večanjem kompleksnosti testnih scenarijev pa se rezultati takšnega razpoznavanja govora bistveno poslabšajo.

V nadaljevanju članka bomo najprej predstavili jezikovne vire, ki smo jih uporabili pri izgradnji sistema avtomatskega razpoznavanja govora. V tretjem poglavju bo sledila kratka predstavitev izdelave akustičnih in jezikovnih modelov eksperimentalnega sistema. Rezultate in analizo vrednotenja razpoznavanja govora bomo predstavili v četrtem poglavju. Zaključek in smernice za nadaljnje delo bomo podali v petem poglavju.

[1] Raziskovalno delo je bilo delno sofinancirano s strani ARRS po pogodbi št. P2-0069.

2. Jezikovni viri

Jezikovni viri predstavljajo eno izmed ključnih točk pri izdelavi razpoznavalnika tekočega govora in igrajo še posebej pomembno vlogo pri jezikih z manjšim številom govorcev, kamor lahko prištevamo tudi slovenski jezik.

2.1. Govorna baza BNSI Broadcast News

Kot osnovo za izvedbo eksperimentov smo uporabili slovensko govorno bazo BNSI Broadcast News (Žgank et al., 2004/1), v enakem obsegu kot za zadnje razviti sistem razpoznavanja tekočega govora (Žgank & Sepesy Maučec, 2010). Baza je na voljo pri organizaciji ELRA/ELDA (ELRA, 2014). Akustični modeli naučeni na takšni bazi so služili kot osnova za nadaljnje eksperimente.

Za izvedbo analize vpliva velikosti učnega korpusa smo uporabili dva dodatna vira transkribiranih posnetkov. Prvi vir so bili preostali posnetki iz govorne baze BNSI Broadcast News, katerih obseg je bil 8 ur. Glede na tip oddaj so se ti posnetki ujemali z do sedaj uporabljenimi posnetki iz govorne baze BNSI Broadcast News. Kot drugi dodatni vir smo uporabili transkribirane posnetke iz interne baze televizijskih oddaj, ki jih bomo v nadaljevanju označevali z IETK-TV. Časovno obdobje teh dodatnih posnetkov ustreza obdobju posnetkov, ki so vključeni v govorno bazo BNSI Broadcast News. Za segmentacijo, označevanje in transkribiranje smo uporabili enake postopke kot za izdelavo baze BNSI Broadcast News (Žgank et al., 2004/1). Oddaje v tem dodatnem delu učne baze po tipu delno odstopajo od oddaj iz govorne baze BNSI Broadcast News, saj posnetki vsebujejo tudi različne intervjuje in omizja, ter s tem pokrivajo bistveno širši spekter televizijskih vsebin in načinov govora. Iz govorne baze IETK-TV smo kot drugi vir učnega materiala uporabili 29 ur dodatnih posnetkov. Baza BNSI Broadcast News v nasprotju z bazo IETK-TV vsebuje samo večerne in nočne dnevno-informativne oddaje, kjer je v večji meri prisoten bran kot spontan govor (Schwartz et al., 1997). Primerjavo lastnosti med govorno bazo BNSI Broadcast News in govorno bazo IETK-TV podajamo v tabeli 1.

Lastnost	***BNSI Broadcast News***	***IETK-TV***
dolžina	*29,86 ure*	28,97 ure
število oddaj	*34*	30
število tipov oddaj	*2*	7
število govorcev	*2073*	784
delež spontanega govora	*30,74%*	68,37%

Tabela 1: Primerjava lastnosti učnega nabora govorne baze BNSI Broadcast News in IETK-TV.

Opravljena primerjava kaže, da vsebuje govorna baza IETK-TV bistveno večji delež spontanega govora pri hkrati manjšem številu različnih govorcev, kar je posledica vključitve različnih intervjujev in omizij v nabor oddaj.

Da bi lahko pravilno ovrednotili vpliv količine posnetkov v učni govorni bazi, smo kot izhodišče vzeli učni govorni korpus iz predhodnih eksperimentov, nato pa smo iz dodatnega materiala korakoma dodajali nove posnetke za učenje akustičnih modelov, in sicer v obsegu: 25%, 50%, 75% in 100%. Na takšen način bomo v nadaljevanju tudi označevali pripadajoče akustične modele.

2.2. Tekstovne baze

Za učenje jezikovnih modelov smo obstoječim besedilnim korpusom v dveh korakih dodali še korpus FidaPLUS.

Korpus FidaPLUS (Arhar & Gorjanc, 2007) je referenčni korpus slovenskega pisanega jezika, ki obsega 621 mio. besed in vsebuje besedila iz tekstovnih virov iz obdobja 1990-2006. Večji del korpusa predstavljajo časopisni in revijalni članki ter knjige. Nekaj besedil izvira tudi iz Spleta. Korpus je lematiziran in označen z morfosintaktičnimi značkami, ki pa v pričujočem članku niso bile uporabljene.

Obstoječa korpusa BNSI-Speech in BNSI-Text sta korpusa govorjenega jezika, obstoječi korpus Večer in novo dodani korpus FidaPLUS pa sta korpusa pisanega jezika. Med govorjenim (predvsem spontano govorjenim) in pisanim jezikom je velika razlika (Stouten et al., 2006). Številnih pojavov, ki so značilni za govor (npr. krajše izjave, uporaba mašil, ponavljanje, napačni starti ipd.), v pisanem jeziku ne zasledimo (Žgank et al., 2008). Ker je FidaPLUS po obsegu neprimerno večja od korpusov govorjenega jezika, bi se z enostavnim razširjanjem učnega korpusa lastnosti govorjenega jezika izgubile. Uravnotežen vpliv različnih jezikovnih virov smo dosegli z linearno interpolacijo na korpusu BNSI-Devel. BNSI-Devel korpus je po strukturi enak korpusu BNSI-Speech in obsega 4 oddaje. Poskrbeli smo, da med korpusi BNSI-Speech, BNSI-Devel in BNSI-Eval (ki je namenjen vrednotenju) ni vsebinskega prekrivanja.

3. Razpoznavalnik govora UMB BN 2014

Sistemi za razpoznavanje tekočega govora so še vedno eni izmed najkompleksnejših na področju govornih tehnologij. V eksperimentih smo kot izhodišče uporabili konfiguracijo iz predhodnih eksperimentov (Žgank & Sepesy Maučec, 2010), ki smo ji delno spremenili modul za izločanje značilk, kjer smo postopku mel-kepstralnih koeficientov dodali normalizacijo srednjih vrednosti kepstra ter normalizacijo energije signala. Za vpeljavo normalizacije smo se odločili zato, ker smo razširili nabor oddaj v učni bazi, ki so imele med seboj delno različne akustične značilnosti.

3.1. Izdelava akustičnih modelov

Osnovo razpoznavalnika tekočega govora predstavljajo zvezni prikriti modeli Markova (HMM) s tristanjsko levo-desno topologijo ter Gaussovimi porazdelitvami funkcije verjetnosti. Kot osnovno enoto akustičnih modelov smo uporabili grafem, ki je že v predhodnih eksperimentih (Žgank et al., 2008; Žgank & Kačič, 2005/1) pokazal učinkovitost delovanja.

Za učenje akustičnih modelov smo uporabili postopek, predstavljen v (Žgank & Sepesy Maučec, 2010). Osnovne značilnosti tako razvitega sistema razpoznavanja govora so podane v tabeli 2.

UMB BN ASR	
Izloč. značilk	MFCC z normalizacijo
Karakteristike značilk	Okno 25 ms, korak 10 ms, MFCC 12 koef., energija, 1. in 2. odvod, 26 filtrov, normalizacija kepstra in energije
Akustični model	Medbesedni (Odell, 1995) trigrafemi
Kompleksnost AM	utežena vsota 16 Gaussovih porazdelitev verjetnosti na stanje
Združevanje AM	odločitveno drevo na osnovi grafemskih razredov (Žgank et al., 2005/2)
Jezik. modeli	Interpolirani trigrami
Vel. slovarja	64.000 besed

Tabela 2: Značilnosti razpoznavalnika tekočega govora.

Sistem, predstavljen v tabeli 2, je služil kot izhodiščni sistem za vrednotenje razpoznavanja govora. Nato smo štirikrat v celoti ponovili postopek učenja akustičnih modelov, vsakič z večjim naborom učnih posnetkov, predstavljenim v poglavju 2.1. Kompleksnost akustičnih modelov smo kontrolirali s konstantno nastavitvijo praga povečanja logaritemske verjetnosti v postopku združevanja z odločitvenim drevesom. Za vrednotenje razpoznavalnika govora smo tako imeli na voljo 5 različnih naborov akustičnih modelov. Razpoznavanje govora smo vedno izvedli z identičnim naborom parametrov dekodirnika.

3.2. Izdelava jezikovnih modelov

Za gradnjo jezikovnih modelov smo uporabili orodje SRI Language Modeling Toolkit (Stolcke, 2002). Na osnovi besedilnih korpusov smo zgradili dva trigramska jezikovna modela. Oba sta bila sestavljena iz štirih komponent: prvo komponento smo zgradili na korpusu BNSI-Speech, drugo na korpusu BNSI-Text, tretjo na korpusu Večer in četrto na korpusu FidaPLUS. V prvem modelu (JMx1) smo uporabili polovico FidePLUS, v drugem pa celo (JMx2). Dokumente za prvi model smo izbirali naključno. V obeh modelih smo v prvih treh komponentah ohranili vse bigrame in trigrame iz učnih korpusov, v četrti komponenti pa smo izločili vse n-grame s frekvenco 1.

Slovar je obsegal 64.000 besed. Vseboval je vse besede korpusov BNSI-Speech in BNSI-Text. Do velikosti 64.000 smo ga dopolnili z najpogostejšimi besedami iz korpusa Večer.

Uporabili smo Good-Turingovo glajenje in sestopanje po Katz-u. Interpolacijske koeficiente komponent smo določili tako, da smo minimizirali perpleksnost jezikovnega modela na korpusu BNSI-Devel. Interpolacijski koeficienti po komponentah za oba modela so predstavljeni v tabeli 3. Utež četrte komponente se je po podvojitvi velikosti zmanjšala.

Komponenta	JMx1	JMx2
BNSI-Speech	0,18	0,18
BNSI-Text	0,23	0,24
Večer	0,11	0,12
FidaPLUS	0,48	0,46

Tabela 3: Koeficienti λ komponent v interpoliranih trigramskih modelih.

Perpleksnost prvega modela na BNSI-Eval je znašala 258, drugega modela pa 246. Delež besed izven slovarja (OOV) je bil 4,22%. Besed, ki so bile izven slovarja, nismo posebej modelirali na nivoju akustičnega modela. Podvojitev velikosti četrte komponente je perpleksnost izboljšala le za 4,56%. Pri tem se je število bigramov v modelu povečalo za 24%, trigramov pa kar za 45%.

4. Rezultati eksperimentov

Vrednotenje sistema za razpoznavanje govora smo izvedli z namenskim evalvacijskim naborom baze BNSI Broadcast News, ki vsebuje 4 različne dnevno-informativne oddaje s 1898 stavki. Rezultate razpoznavanja govora smo podali v odstotku pravilno razpoznanih besed, kjer smo upoštevali tudi vrinjene oz. izbrisane besede, ki posledično dodatno poslabšajo rezultat.

Akustični modeli	***Pravilno razpoznane besede (%)***
+00%	*72,44*
+25%	*73,10*
+50%	*72,87*
+75%	*73,00*
+100%	*73,19*

Tabela 4: Rezultati razpoznavanja govora za večanje učne baze akustičnih modelov z jezikovnim modelom JMx1.

Izhodiščni sistem razpoznavanja govora, ki je uporabljal osnovne akustične modele (+00%), ter osnovni jezikovni model (JMx1) je dosegel 72,44% pravilnost razpoznavanja besed na evalvacijskem naboru posnetkov baze BNSI Broadcast News. Doseženi rezultat je primerljiv z razpoznavalniki govora za slovenski jezik, ki uporabljajo podobno kompleksno zasnovo (Žgank in Sepesy Maučec, 2010).

V naslednjem koraku smo z uporabo osnovnega jezikovnega modela JMx1 ovrednotili različne akustične modele (+25% … +100%), naučene na večjem naboru učnih posnetkov. Rezultati razpoznavanja tekočega govora so se izboljšali na 72,87% (+50%) do največ 73,19% z akustičnimi modeli +100%. Tako smo s povečanjem učne baze posnetkov za približno dvakrat dosegli izboljšanje rezultatov razpoznavanja govora za največ 0,75% absolutno. Glede na količino dodatnega učnega materiala, ter potrebni finančni in časovni vložek za njegovo pripravo, lahko trdimo, da je izboljšanje rezultatov razpoznavanja govora relativno majhno.

Akustični modeli	***JMx1, Pravilnost razpoznavnja besed (%)***	***JMx2, Pravilnost razpoznavnja besed (%)***
+00%	*72,44*	72,53
+25%	*73,10*	72,97
+50%	*72,87*	73,02
+75%	*73,00*	73,15
+100%	*73,19*	73,30

Tabela 5: Rezultati razpoznavanja govora z jezikovnim modelom naučenim na večjem korpusu.

V zadnjem koraku vrednotenja smo analizirali, kako na rezultate razpoznavanja govora vpliva povečanje

učnega korpusa jezikovnega modela za dvakrat (JMx2). Rezultati so predstavljeni v tabeli 5.

V izhodišču smo s povečanjem učnega korpusa za pripravo jezikovnega modela za dvakratnik (JMx2) uspeli doseči rezultat razpoznavanja govora 72,53%. Izboljšanje rezultata je znašalo 0,09% absolutno, kar predstavlja minimalno razliko, še posebej če upoštevamo, kako veliko povečanje besedilnega korpusa in posledično jezikovnega modela je bilo potrebno za dosego tega rezultata.

Podobna minimalna izboljšanja rezultatov z novim jezikovnim modelom JMx2 smo dosegli tudi v kombinaciji z različnimi akustičnimi modeli naučenimi na povečani govorni bazi. Edina izjema so bili akustični modeli +25%, kjer je z večjim jezikovnim modelom JMx2 prišlo celo do rahlega poslabšanja rezultatov, in sicer za 0,13% absolutno.

V kombinaciji akustičnih modelov +100% in jezikovnega modela JMx2, kjer smo obakrat uporabili največje razpoložljive govorne in jezikovne vire, smo dosegli skupno najboljši rezultat razpoznavanja govora 73,30%, kar sicer predstavlja 0,86% absolutno izboljšanje, vendar je pri tem potrebno upoštevati, koliko večji viri so bili potrebni za dosego takšnega rezultata.

Na dobljenih rezultatih smo izvedli test statistične značilnosti, pri čemer smo za mejo statistične značilnosti izbrali vrednost α=0,05. Izkazalo se je, da je statistično značilna (0,012) le razlika pri povečanju govorne baze (+100%), kadar uporabljamo jezikovni model JMx2. Vse ostale primerjave so pokazale, da izboljšanja niso statistično značilna za izbrano mejo. Ti rezultati dodatno potrjujejo našo domnevo, da samo s povečevanjem učnega materiala ne moremo doseči bistvenih izboljšav razpoznavanja slovenskega tekočega govora.

Če hipotetično predpostavimo, da bi se ohranil takšen trend povečevanja pravilnosti razpoznavanja besed z večanjem obsega učnih virov (v kar avtorji sicer dvomimo), bi za doseganje pravilnosti razpoznavanja besed vsaj 90%, po oceni potrebovali tematsko ustrezne vire v obsegu več kot 1100 ur transkribiranega govora in jezikovne vire z več kot 12,1 giga besed. Takšen predpostavljen obseg virov za večkratni faktor presega vse do sedaj ustvarjene vire za slovenski jezik.

5. Zaključek

V članku smo poskusili odgovoriti na vprašanje, kako pomembna je velikost govornih in jezikovnih virov za izboljšanje rezultatov razpoznavalnika govora za slovenski jezik. Analiza eksperimentov je pokazala, da je s povečanjem virov sicer možno doseči minimalno izboljšanje rezultatov, vendar so za dosega tega cilja potrebna velika vlaganja v izdelavo virov. Pokazalo se je, da je pri tem izrednega pomena tudi ujemanje virov v žanru oz. domeni.

Na osnovi doseženih rezultatov lahko z dokaj veliko verjetnostjo predpostavimo, da za visoko pregibni slovenski jezik ne zadostuje samo večanje obsega govornih in jezikovnih virov, temveč da je hkrati potrebno tudi nadaljevati z raziskovalnim delom na področju algoritmov, ki bi ustrezno naslavljali specifične lastnosti visoko pregibnega slovenskega jezika.

Zahvala

Zahvaljujemo se avtorjem besedilnega korpusa FidaPLUS, ki so nam omogočili njegovo uporabo za jezikovno modeliranje avtomatskega razpoznavalnika govora.

6. Literatura

Arhar, Š., Gorjanc, V., (2007). Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. Jezik in slovstvo 52/2., 95--110.

Dobrišek, S., Mihelič, F., (2010). Zmanjševanje odvečnosti končnih pretvornikov za učinkovito gradnjo razpoznavalnikov slovenskega govora z velikim besednjakom. *Jezikovne tehnologije*, Ljubljana, Slovenija.

ELRA (2014). *BNSI Catalog Reference : S0275*: www.elra.info.

Odell, J.J., (1995). *The Use of Context in Large Vocabulary Speech Recognition*. Doktorska disertacija, Univerza v Cambridgeu, Velika Britanija.

Rotovnik, T., Sepesy Maučec, M., Kačič, Z. (2007). Large vocabulary continuous speech recognition of an inflected language using stems and endings. *Speech communication*, 2007, vol. 49, iss. 6, 437--452.

Schwartz, R., Jin, H., Kubala, F., Matsoukas, S., (1997). Modeling those F-Conditions - or not. *Proc. DARPA Speech Recognition Workshop*, Chantilly, ZDA.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. International Conference on Speech and Language Processing, II: 901--904.

Stouten, F., Duchateau, J., Martens, J.-P., Wambacq, P., (2006). Coping With Disfluencies In Spontaneous Speech Recognition:Acoustic Detection And Linguistic Context Manipulation, Speech Communication vol. 48, issue 11, 1590--1606.

Žgank, A., Rotovnik, T., Sepesy Maučec, M., Verdonik, D., Kitak, J., Vlaj, D., Hozjan, V., Kačič, Z., Horvat, B., (2004/1). Acquisition and annotation of Slovenian broadcast news database. *Fourth international conference on language resources and evaluation, LREC 2004*, Lizbona, Portugalska.

Žgank, A., Kačič, Z., Vicsi, K., Szaszak, G., Diehl, F., Juhar, J., Lihan, S., (2004/2). Crosslingual transfer of source acoustic models to two different target languages. *Robust2004 : COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*, Norwich, Velika Britanija.

Žgank, A., Kačič, Z., (2005/1). Primerjava treh tipov akustičnih osnovnih enot razpoznavalnika slovenskega govora. *Elektrotehniški vestnik*, 2005, Ljubljana, Slovenija.

Žgank, A., Horvat, B., Kačič, Z., (2005/2). Data-driven generation of phonetic broad classes, based on phoneme confusion matrix similarity. *Speech Communication*, vol. 47, issue 3, 379--393, november 2005.

Žgank, A., Kos, M., Kotnik, B., Sepesy Maučec, M., Rotovnik, T., Kačič, Z. (2008). Nadgradnja sistema za razpoznavanje slovenskega tekočega govora UMB Broadcast news. *Jezikovne tehnologije 2008*, Ljubljana, Slovenija.

Žgank, A., Sepesy Maučec, M., (2010). Nadgradnja sistema Razpoznavalnik tekočega govora UMB Broadcast News 2010: nadgradnja akustičnih in jezikovnih modelov. *Jezikovne tehnologije 2010*, Ljubljana, Slovenija.

Vprašanja zapisovanja govora v govornem korpusu Gos

Darinka Verdonik

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko
Smetanova 17, SI-2000 Maribor
darinka.verdonik@um.si

Povzetek

Prispevek obravnava vprašanja, povezana z morebitno nadgradnjo referenčnega govornega korpusa slovenščine Gos, s poudarkom na nekaterih težavnejših vprašanjih zapisovanja govora. Morebitna nadgradnja v smeri skupne platforme z akustično bazo, potrebno za razvoj razpoznavanja tekočega govora, pri vprašanjih zapisovanja govora predvideva nadaljnji zapis po dvotirnem sistemu pogovornega in standardiziranega zapisa, vzpostavljenem v korpusu Gos. Vendar obstoječe gradivo Gosa kaže nekatere nedoslednosti in odprta vprašanja, ki bi jih bilo treba pri nadgradnji odpraviti. Izpostavimo štiri: vprašanje zapisovanja dvoustničnega 'U' in člena 'ta' v pogovornem zapisu, vprašanje zapisovanja neverbalnih in polverbalnih glasov ter vprašanje standardizacije nestandardnih polnopomenskih izrazov.

The questions of speech transcription in the speech corpus GOS

In this paper we discuss issues related with eventual upgrade of the reference speech corpus GOS, with special attention to questions concerning the speech transcription. It is likely that the upgrade of the GOS corpus will be joined with efforts to provide new acoustic speech database for continuous speech recognition. Nevertheless, the transcription of speech should follow the two-level transcription system (pronunciation-based and standardized transcription) specified in the GOS corpus. However, the existing transcriptions of the GOS show some inconsistencies and open questions that need to be discussed before the upgrade. In this paper, we discuss four such issues: the transcription of the sonorant phoneme U and the particle 'ta' in the pronunciation-based transcription, the transcription of non-verbal and semi-verbal sounds in the pronunciation-based and standardized transcription, and the transcription of non-standard lexical items in the standardized transcription.

1. Uvod

Konec leta 2010 je bil v slovenskem prostoru javnosti predstavljen prvi poskusni referenčni korpus govorjene slovenščine – Gos (Verdonik, Zwitter Vitez, 2011; Verdonik idr., 2013). Poskusni pravimo zato, ker pravi referenčni (pisni) korpusi obsegajo po več 100 milijonov besed, govorni le po nekaj milijonov, Gos 1 milijon. Kot s(m)o zapisali avtorji ob njegovi objavi na spletu, zato vsi upamo, da bo v prihodnosti še rasel, in ta prispevek izhaja iz tega upanja.

Toda zdi se, da vsakdanji govorjeni jezik ne požanje prav veliko zanimanja jezikoslovcev; niti toliko, kot ga kažejo sami govorci, ki se ob pomanjkanju aktivnosti na strani stroke v posameznih iniciativah lotijo tudi njegovega opisovanja (gl. npr. http://www.pokazijezik.si/ ali http://razvezanijezik.org/). Na drugi, tehnološki strani jezikovne tehnologije (z vmesnimi padci) vsake toliko »udarijo« z govornimi tehnologijami – na primer s govornim sistemom dialoga, kot je Siri na iPadu, ali s strojnim prevajanjem govora, kot je pred nekaj meseci Microsoft s prevajalnikom govora za Skype. Seveda (se bomo sploh kdaj znebili tega seveda?) pa so ti sistemi narejeni za velike tuje jezike, slovenščine ne pokrivajo, in tako je tudi na tem področju v slovenskem okolju zanimanje za govorjeni jezik prepuščeno le redkim posameznikom, ki se s temi tehnologijami ukvarjajo.

Ob tem, da prav velikega zanimanja za vprašanja vsakdanjega govorjenega jezika v stroki ne zaznamo, pa se hkrati čudimo, da tako pogost in vseprisoten pojav ostaja tako neraziskan in nezanimiv za raziskovalce (in s tem mislimo predvsem jezikoslovce). Morda pa je glavni razlog samo težavnost zbiranja in urejanja gradiva ter majhnost obstoječega govornega korpusa ... in s to mislijo smo se lotili njegove analize in načrtov za prihodnjo rast, optimistično odločeni, da se bo slednja prej ali slej zgodila.

2. Nadaljnja rast v smeri večje podpore razvoju tehnologij

Korpus Gos je bil načrtovan predvsem kot reprezentativni korpus za jezikoslovne raziskave. Kljub temu je vsaj del korpusa, zlasti tisti, ki predstavlja javni diskurz (to je nekaj 10 ur govora), mogoče uporabiti tudi kot akustično bazo za razvoj razpoznavanja tekočega govora,[1] čeprav ni v celoti prilagojen tovrstni uporabi.

Ob rasti akustičnih baz za razpoznavanje tekočega govora v mednarodnem prostoru in primerjavi z razpoložljivimi bazami za slovenščino (med temi predvsem BNSI Broadcast News – Žgank idr. 2004; slovenska govorna baza Broadcast News – Žibert, Mihelič, 2004; SloParl – Žgank idr., 2006) pa postaja več kot očitno, da je velika ovira za nadaljnji razvoj tovrstne tehnologije za slovenščino prav pomanjkanje ustreznega obsega akustičnih baz, ki se meri v tujini v nekaj sto urah, za slovenščino zaenkrat samo v nekaj deset urah. Na drugi strani je za obstoječi reprezentativni govorni korpus Gos prav tako treba načrtovati velik preskok v obsegu, v mednarodnem prostoru postaja referenčni obseg za primerljive vire ca. 10 mio. besed.

Aktualni akcijski načrt za jezikovno opremljenost (Dobrovoljc idr., 2014) predvideva v nadaljnjih načrtih na področju govornih korpusnih in akustičnih virov skupno platformo tako za nadgradnjo referenčnega govornega korpusa Gos kot akustične baze za razpoznavanje tekočega govora. V tej smeri je zastavljena tudi primerjalna analiza korpusa Gos z bazo BNSI Broadcast News kot predstavnikom akustične baze, izdelane za potrebe razpoznavanja govora, objavljena letos na

[1] Možnost uporabe gradiv za razpoznavanje govora velja samo za tisti del gradiv, ki predstavljajo javni diskurz. Posnetke in transkripcije je v ta namen mogoče pridobiti prek konzorcija CLARIN.SI.

konferenci LREC (Žgank idr., 2014). Povzemimo na kratko rezultate te analize.

Razlike med obema viroma so bile ugotovljene na petih ravneh:

(1) (Delno) različna je vsebina enega in drugega vira: reprezentativni govorni korpus zajema vzorce iz vseh najpogostejših tipov govorne interakcije, kar vključuje cel spekter od medijskega diskurza na eni do zasebnih pogovorov na drugi strani, akustična baza pa se osredotoča na zajemanje posnetkov s področij, kjer je najverjetnejša aplikacija uporabe, to so pri BNSI televizijski diskurzi (možnost aplikacije pri avtomatskem podnaslavljanju in prevajanju), sicer pa še drugi bolj kot ne formalni javni govorni nastopi (npr. parlamentarne seje, razna javna predavanja in predstavitve ipd., kjer je možnost aplikacije za avtomatsko izdelovanje dobesednih zapisov), za gluhe in naglušne pa je zanimivo področje aplikacije izobraževanje.

(2) Na akustični ravni je pri korpusu Gos ugotovljen veliko širši spekter različnih akustičnih okolij kot v bazi BNSI, hkrati pa zelo skopa označenost akustičnih okolij in slaba kvaliteta zajema avdio signala. Smernice za skupen jezikovni vir so zato predvidene v smeri večje kvalitete zajema avdio signala in bolj natančnega označevanja akustičnega okolja (zlasti tehnologije zajema signala in akustičnega ozadja, kot je npr. govor, hrup, glasba ...).

(3) Na ravni segmentiranja govora so ugotovljene razlike v načinu določanja segmentov, saj je pri akustični bazi veliko pozornosti usmerjene v ločevanje hkratnega govora in premorov, tudi iskanje mej med segmenti sledi v prvi vrsti ustrezno dolgim premorom v govoru, medtem ko v govornem korpusu označevanje hkratnega govora ni natančno, segmenti pa sledijo v prvi vrsti smiselno zaokroženim izjavam. Skupne smernice so predvidene v smeri, ki jo zastavi korpus Gos, z dodatkom, da se bolj kot v obstoječi praksi transkribiranja v korpusu Gos sledi načelu čim krajših segmentov, zlasti tistih, kjer se pojavlja hkratni govor, in da se bolj podrobno kot doslej označujejo premori v govoru.

(4) Naslednja razlika je opredeljena kot raven označevanja akustičnih dogodkov, to so razni vdihi, izdihi, tleski z jezikom ipd. oz. negovorni zvoki (npr. zvonjenje). Ti dogodki morajo biti v akustični bazi natančneje označeni, v govornem korpusu pa so bili označeni le pragmatično pomembni. Skupne smernice so predvidene v smeri bolj podrobnega označevanja.

(5) Različni praksi sta tudi na področju zapisovanja govora. V akustični bazi BNSI sledi zapis pravopisnemu standardu, če izgovorjava opazno odstopa od predvidene standardne, pa so dodane posebne oznake k takim besedam. V korpusu Gos pa je bil razvit dvotirni sistem zapisovanja govora, kar je bil učinkovit način za obvladovanje številnih izgovornih različic, ki se pojavljajo zlasti pogosto v nejavnem diskurzu. Skupne smernice predvidevajo nadaljevanje dvotirnega zapisovanja, in tega bomo prav zato v tem prispevku nekoliko podrobneje analizirali. Zapis govora je namreč do neke mere vedno interpretacija tistega, kar slišimo. Pri tem se srečujemo z mnogimi vprašanji, kako oblikovati načela zapisovanja, da bomo ohranili vse pomembne jezikovne prvine in hkrati omogočili čim večjo mero avtomatskega prepoznavanja posameznih prvin. Nekaterim od teh vprašanj, ki so se odprla ob uporabi korpusa, se bomo posvetili v nadaljevanju. Zagotovo pa to niso vsa vprašanja zapisovanja govora in zaželeno bi bilo, da se v prihodnosti vedno znova kritično ozremo nazaj.

3. Načela zapisovanja govora v Gosu

Zapisovanje govora v Gosu je bilo zasnovano po dvotirnem sistemu, ki je bolj kot ne unikaten tudi v svetovnem merilu (gl. npr. Verdonik idr., 2013). V specifikacijah korpusa (http://www.korpus-gos.net/Content/Static/Nacela_transkribiranja_in_oznacevanja_posnetkov_v_referencnem_govornem%20korpusu_slovenscine.pdf) je dvotirni zapis utemeljen in opisan takole:

»Pri zapisu govora se je hitro pokazalo, da nekaterih ciljev (hitro in enostavno transkribiranje, dejanska podoba diskurza, avtomatsko iskanje po besednih oblikah z enako oblikoslovno in semantično vlogo, a različnimi glasovnimi podobami) ni mogoče rešiti z eno samo rešitvijo.

Zato smo ustvarili dva nivoja zapisa govora: na prvem nivoju zapisa, ki ga imenujemo 'pogovorni zapis', zapišemo besede sicer ortografsko (ne fonetično!), vendar tako, kot so izgovorjene; na drugem nivoju, ki ga imenujemo 'knjižni zapis' (kasneje spremenjeno v 'standardizirani zapis', op. a.), pa 'poknjižimo' zapis na tak način, da različnim variantam neke besedne oblike (npr. *mam, jemam*) pripišemo krovno knjižno obliko (npr. *imam*).

Tako s prvim nivojem omogočimo dober vpogled v besedje in oblike govorjenega jezika, z drugim nivojem pa razširimo iskalne možnosti ter omogočimo uspešnejše nadaljnje avtomatsko označevanje besedil.«

Za ilustracijo, kako sta oba nivoja zapisa realizirana v praksi, navajamo v nadaljevanju primer iz Gosa:

Pogovorni: *ne sej tak eee tak ko si razložla men mislim veš kak je s temi sanjami ne*

Standardizirani: *ne saj tako eee tako kot si razložila meni mislim veš kako je s temi sanjami ne*

Vseh podrobnosti enega in drugega nivoja zapisa tukaj ne bomo obravnavali, pojasnjene so na spletni strani Gosa (www.korpus-gos.net) v priloženih specifikacijah in v monografiji (Verdonik, Zwitter Vitez, 2011).

4. Zapisi govora v Gosu v številkah

Gos vsebuje 1,035.101 besedo v standardiziranem zapisu. Tabela 1 prikazuje, koliko od teh besed je različnic na nivoju pogovornega in standardiziranega zapisa ter leme.

Tabela 1: Število različnic v Gosu

Nivo	**Št. različnic**
pogovorni zapis	82.648
standardizirani zapis	62.578
lema	31.294

Vsaki besedi v pogovornem zapisu je pripisana ena (izjemoma pa lahko tudi dve ali več) beseda v standardiziranem zapisu. Tabela 2 prikazuje, koliko je vseh tovrstnih parov različnic, koliko je identičnih in koliko neidentičnih ter nakaže strmo padanje frekvenc pojavitve pri neidentičnih parih. Strmo padanje je verjetno delno posledica majhnosti korpusa, je pa tovrstna krivulja frekvenc v jeziku nasploh značilna.

Tabela 2: Pari besed pogovorni – standardizirani zapis

Pari pogovorni – standardizirani zapis	**Število (% vseh parov)**
vseh parov	82.648
identičnih parov	54.822 (66 %)
neidentičnih parov	27.826 (34 %)
neidentičnih parov, ki se pojavijo več kot 5-krat	3.391 (4 %)
neidentičnih parov, ki se pojavijo več kot 100-krat	210 (0,25 %)
neidentičnih parov, ki se pojavijo več kot 1000-krat	18 (0,02 %)

Deset najpogostejših neidentičnih parov je naslednjih, po pričakovanju funkcijskih besed, saj so te v jeziku najpogosteje rabljene:

Po.:	St.:	
tud	*tudi*	*3571*
jz	*jaz*	*3460*
sej	*saj*	*3399*
al	*ali*	*3251*
zdej	*zdaj*	*3036*
tko	*tako*	*2820*
tak	*tako*	*2667*
blo	*bilo*	*2263*
sam	*samo*	*1699*
sn	*sem*	*1620*

5. Nekatera težavnejša vprašanja zapisovanja govora

Tukaj ne bomo obravnavali vseh načel zapisovanja govora v korpusu Gos, ampak samo nekatera težavnejša vprašanja. Prvi dve se nanašata na pogovorni zapis, kjer ponekod opazimo nedoslednosti, tretje na pogovorni in standardizirani zapis ter četrto na standardizirani zapis.

5.1. Dvoustnični 'U'

Načelo pogovornega zapisa številka 3 v specifikacijah transkribiranja za korpus Gos pravi: »Dvoustnični v zapisujemo s črko 'v' (*prov, nav, navm, odpravt, davn*...) oz. tudi z 'l', če tako izhaja iz knjižne norme (*kosil* (v pomenu *kosilo*), *mel* (v pomenu *imel*)). Če je u samoglasniški, ga pišemo s črko 'u' (*pršu, vidu*...).«

Zdi se, da tovrstno načelo govorcem slovenščine vseeno ni popolnoma domače, ko morajo zapisovati besede govorjene slovenščine, ki še nimajo ustaljenega »standarda« zapisovanja, in sicer se marsikje namesto predvidenega zapisa z 'v' ali 'l' vrine zapis z 'u' – npr. *laufati*, *šlauf* ali *genau* se v zapisu z 'u' pojavljajo celo v Besedišču in tudi po korpusu Gigafida močno prevladuje različica z 'u', čeprav bi po zgornjem pravilu pisali *lavfati, šlavf, genav*. Podobno so dvojnice lahko pri medmetih, npr. *au* in *av* (po SSKJ).

V zvezi s tem se pojavlja tudi nekaj več nedoslednosti v pogovornem zapisu korpusa Gos, kjer najdemo po večkrat tudi pogovorne zapise tipa *mau* (*malo*), *biu* (*bil*), *šou* (*šel*), *dou* (*dol*), *prou* (*prav*), *dau* (*da bo*), *nou* (*ne bo*) itd., namesto predvidenega zapisa s črko v/l. Kljub temu pa je večinsko zapis z v/l v tovrstnih vlogah prevladujoč in zdi se, da bi bilo spreminjanje načela v zapis z 'u' še bolj problematično: potem bi namreč besede, ki v glasovni podobi sledijo standardu, še vedno pisali z 'v' ali 'l', npr. *imel*, in kontrast z *meu* namesto *mel* bi verjetno vnesel še več zmede in nedoslednosti. Edina sprejemljiva sprememba tega pravila bi zato bila, da se vodi seznam besed ali oblik, za katere lahko po pisnih korpusih sledimo tendenci po pisanju s črko 'u' v teh položajih, ostale pa se še naprej pišejo z 'v' oz. 'l'. Je pa vprašanje, ali ni tako pravilo še bolj problematično s stališča doslednosti zapisovanja kot obstoječe uniformno vodilo.

5.2. Člen 'ta'

Določila, ki bi posebej omenjalo pisanje člena 'ta' v tipu 'ta rdeči' (kjer je 'ta' nenaglašen in izgovorjen skupaj s sledečim pridevnikom), v specifikacijah transkribiranja ni bilo, iz korpusa pa vidimo, da se je sledilo praksi, da se člen piše kot samostojna beseda. Ob tem pa na nivoju pogovornega zapisa (kot posamezne lapsuse pa posledično tudi na ravni standardiziranega zapisa) vseeno občasno zasledimo stični zapis, zelo pogosto za zvezo *ta mali/ta mala,* npr. *tamal, tamav*, *tamalo, tamali, tamalima, tamavga, tamalga*, poleg te pa bolj kot ne posamično še za zveze *taprav/tapravo, tapravga* (*ta pravi*), *tazaden* (*ta zadnji*), *tamladi* (*ta mladi*), *taprv* (*ta prvi*), *tazadno* (*ta zadnjo*) itd.

Medtem ko je na nivoju standardiziranega zapisa res najbolj praktično in smiselno nestično pisanje, zlasti z vidika kasnejšega oblikoslovnega označevanja, izdelave besednih seznamov in iskanja po besedilu, pa bi veljalo še enkrat razmisliti o možnosti stičnega pisanja v pogovornem zapisu. S tem bi namreč omogočili avtomatsko ločevanje med rabami tipa zaimek + pridevnik (*hvala za ta lep mejl*) in rabami tipa člen + pridevnik (*je bil predračun tak da je šu tist talep lijak ven*), ki jih je mogoče zanesljivo ločevati samo ročno in s pomočjo zvočnega posnetka.

5.3. Neverbalni in polverbalni izrazi

O pisanju neverbalnih in polverbalnih izrazov govori določilo pogovornega zapisa številka pet, ki je (opredeljeno vnaprej, pred začetkom transkribiranja) dokaj skopo: "Podaljšane neleksikalne enote pri iskanju formulacije pišemo s tremi črkami, in sicer: *eee, eem, mmm*... oziroma z nizom črk, ki najbolje ustreza dejanski izgovorjavi."

O pretvorbi teh zapisov v standardizirani zapis je v specifikacijah transkribiranja za korpus Gos določilo: "Onomatopeje, medmete, besedne fragmente in druge glasove, za katere v knjižnem jeziku ni standardnega zapisa, pustimo zapisane tako, kot so bili zapisani v prvotni transkripciji," v monografiji (Verdonik, Zwitter Vitez 2011: 67) pa: "Onomatopeje, medmete, besedne fragmente in druge glasove standardiziramo z enotno krovno obliko, kjer je to mogoče: *jooj, ijoj* > *joj*." Sprememba določila je posledica opažanja, da so v pogovornem zapisu nastajale nedoslednosti pri zapisovanju.

Vseeno pa obstoječa rešitev, da so nekatere glasovno različne realizacije neverbalnih ali poverbalnih glasov vodene pod enotnim krovnim zapisom, ni povsem idealna, saj tukaj večinoma ne moremo govoriti o redukcijah ali glasovnih premenah kot pri bolj verbaliziranih enotah. Za primer: pri *ijoj* ne moremo govoriti o glasovni premeni osnove *joj*.

Neverbalni in polverbalni glasovi so, gledano površinskobesedilno, ena najbolj pogostih in tipičnih

značilnosti govorjenega besedila, ob tem pa povzročajo težave tako lematizaciji kot oblikoslovnemu označevanju, učenemu na pisnih besedilih, in so zato pogosto kar sistematično narobe označeni. Gre torej za vprašanje, ki lahko ima na rezultate iskanja po korpusu precejšen vpliv. Problemu smo se zato podrobneje posvetili: pregledali smo vse tovrstne izraze v Gosu in izdelali predlog natančnejših načel njihovega zapisovanja skupaj s seznamom zapisov za te izraze v obstoječem gradivu korpusa Gos. Čeprav za slovenščino pri ZRC SAZU sicer obstaja slovarček medmetov, ki zajema tudi nekatere polverbalne izraze (http://bos.zrc-sazu.si/cgi_new/medmeti/a01.exe?name=medmeti&expression=*), pa je naš seznam prvi, ki temelji na avtentičnem govorjenem gradivu. Seznam je v prilogi 1 tega prispevka in je (med drugim) pomemben predvsem za uspešnejšo lematizacijo in oblikoslovno označevanje govornega gradiva.

Seznam neverbalnih in polverbalnih izrazov v Gosu smo zbrali tako, da smo ročno pregledali seznam standardiziranih zapisov korpusa Gos in iz njega izločili kandidate črkovnih nizov za tovrstne izraze, nato pa jih preverjali prek Gosovega konkordančnika (www.korpus-gos.net). Po pregledu smo za veliko izrazov predlagali nov, popravljen zapis, ki sledi načelom zapisovanja, kot jih povzemamo spodaj.

Načela zapisovanja izhajajo iz dveh stališč: način zapisa naj bi bil govorcem slovenščine čim bližji, hkrati pa naj bi omogočal največjo možno mero avtomatskega procesiranja teh izrazov v govorjenem besedilu. Načela so:

1. izraze zapišemo raje z eno besedo kot več besedami (npr. *ojoj* namesto *o joj*),
2. kjer ni bistvene razlike v zvočni podobi in funkciji/pomenu, ohranimo enoten zapis za različne rabe (npr. *mhm* bi posamično morda zapisali tudi kot *ehm*, vendar je razmejitev težko objektivno določiti, zato raje ohranjamo vedno *mhm*),
3. izraze zapisujemo prednostno s tremi črkami, tako da se razlikujejo od drugih besed (npr. raje *vaa* kot *va*), razen kjer ni nevarnosti, da bi bil zapis identičen zapisu kakih drugih besed, ali če je drugačen zapis že močno uveljavljen (npr. *eh*),
4. dvoustnični U prednostno pišemo z 'v' (*av, vav*),
5. podaljševanje glasov se ne označuje z več črkami, ampak se ohranja enoten zapis (npr. vedno *jee*, ne *jeee* ali podobno),
6. prednost ima poslovenjen zapis (npr. *jes*, ne *yes*, *okej*, ne *ok* ali *okay*).

Kot izstopajoč neenotni in avtomatsko težko sledljiv zapis v obstoječem gradivu izpostavimo neverbalno glasovno zanikanje. Zasledili smo naslednje različice zapisovanja tega pojava: *n n, m m, a a, e e, nn, aa, mm*. Glasovno le-to dejansko niha od bolj vokalnega, a-jevskega prek polglasniškega do zvočniškega m ali n. Da bi bilo neverbalno glasovno zanikanje avtomatsko sledljivo, bi bil potreben bolj enoten in unikaten zapis. Predlagamo dve različici zapisa: *nn* in *aa*.

Nasproten, sicer redkejši primer je neverbalno glasovno pritrjevanje, za katerega je bil realiziran zapis *mm* – ta je primeren, bi pa bilo dobrodošlo, da z njim niso zapisane še kake druge realizacije, npr. zanikanje (kjer predlagamo *nn*) ali oporni signal (*mmm* oz. *eee*).

Iz zgornjih primerov vidimo, da je lahko transkripcija govora na določenih točkah že močno v vlogi interpretacije funkcij/pomenov izrazov. To se zdi dopustno le izjemoma, ko sicer zelo težko enoumno določimo zapis.

Naš predlog je, da se načelom zapisa, kot so predstavljena zgoraj, sledi že pri pogovornem zapisu. Standardizirani zapis se potem za neverbalne in polverbalne glasove ne bi spreminjal, identičen zapis pa bi dobila tudi lema teh izrazov.

5.4. Nestandardni polnopomenski izrazi

Zadnje vprašanje, ki ga bomo obravnavali, se odpira pri standardizaciji zapisa za nestandardne polnopomenske izraze (s tem mislimo take, ki niso sprejeti v standardni jezik), od katerih imajo mnogi v različnih regijah nekoliko različno glasovno realizacijo. Teh ni toliko, kot bi morda pričakovali, vseeno pa dovolj, da je treba njihov krovni standardizirani zapis bolj natančno določiti. V obstoječih specifikacijah Gosa piše: »Pogovorne besede, ki bi jim težko določili povsem ustrezno knjižno različico, ohranjamo. Pri odločitvah glede zapisa se opiramo na pisne korpuse in druge vire.« (http://www.korpus-gos.net/Content/Static/Navodila_za_standardizacijo_zapisa_govora.pdf) Sledijo primeri, ki dodatno ilustrirajo različne rabe, vendar večinoma razne funkcijske besede, polnopomenske pa le na kratko, in sicer pretežno v naslednjem odstavku: »Ohranimo: a) izposojenke *bek, čuješ, fak, fajrala, ferker, ful, gruntali, hambrt, kafič, kao, kuhla, može, ni mus, ornk, pašeš, plata, pošlihtaš, rajsar, ratati, singl, spedenan, štima, šparati, valjda, ziher, žijaš...*«

Nedoslednosti v zapisu v zvezi s tem problemom smo zasledili bolj kot ne naključno, ob uporabi korpusa in pregledovanju njegovih besednih seznamov. Tako se je na primer v standardiziranem zapisu pojavljalo *fertig* in *fertik*, *frej* in *fraj*, *kafe* in *kofe* ...

Poskus sistematičnega sledenja tem pojavom smo naredili tako, da smo najprej izdelali besedni seznam vseh pojavitev v pogovornem zapisu korpusa Gos, nato pa besedni seznam oblik v korpusu Kres, ki se pojavijo vsaj desetkrat. Nato smo besedni seznam korpusa Gos filtrirali s pomočjo besednega seznama korpusa Kres in ročno pregledali samo tiste pojavitve, ki jih ni bilo na besednem seznamu korpusa Kres. Izločili smo kandidate za podrobnejši pregled ter si zanje izpisali konkordance. Te smo ponovno ročno pregledali. Po pregledu seznamov na črki a in b smo na tak način našli štiri dodatne primere, kar potrjuje, da ne gre za zelo obsežen problem, vseeno pa se ne sme ignorirati. Zadeva namreč leksiko, ki je v pisnih korpusih redko prisotna in lahko na primer pri izdelavi geslovnikov ali analizah ključnih besed izpade iz rezultatov ali je v rezultatih neustrezno rangirana, če ni v različnih realizacijah standardizirana vedno na enoten način. Zato se zdi potrebno, da se pri nadgradnji korpusa Gos na tovrstne primere bolj podrobno opozarja in se jih sistematično vodi.

V glavnem naključno zbrani primeri, ki smo jih sami pregledali, kažejo sledeče:

1. različic ne zasledimo: *kao*, *talam/tala* (za lemo *talati*) ...,
2. različne pogovorne realizacije nestandardnih polnopomenskih izrazov so nedvoumno posledica znanih regionalnih in narečnih glasovnih premen, npr. *šihtu* vs. *šejhti* (za lemo

šiht), *cajt* vs. *cet* (za lemo *cajt*), *pasalo* vs. *pasal* (za lemo *pasati*) ...

Medtem ko v zgornjih primerih neenotnega standardiziranega zapisa ne zasledimo, pa ga v naslednjih primerih:

1. ob redukciji določenega glasu: *luškan/lušno* vs. *luštkan/luštno*, *magari* vs. *magar, glej* vs. *lej* ...; ob tem odločitve niso nujno enostavne, zlasti ko je reducirana oblika zelo pogosta ali dobi nove pomene/funkcije; tako na primer tudi SSKJ obliki *lej* (od *glej*) pripiše posebno geslo, v Gosu pa so v zvezi s tem zanimivi še primeri *kurc* vs. *kurac* ali *dedec* vs. *dec* vs. *ded*; težavnost obravnavanja redukcij se kaže tudi na primeru *čmo* vs. *hočemo* (glej specifikacije standardiziranega zapisa, www.korpus-gos.net) in predstavlja pravi jezikoslovni izziv pri nekaterih funkcijskih besedah, npr. *te* vs. *potem*, *k* v vlogi *ker, ko, ki, kot, kjer, kar, kaj* ...;
2. zaradi premen po zvenečnosti, npr. *fertig* vs. *fertik*, *oreng* vs. *orenk* ...;
3. zaradi premen vokalov, npr. *fraj* vs. *frej*, *kafe* vs. *kofe* ...

Medtem ko pri zgornjih primerih prepoznavamo pomanjkanje enotne standardizirane oblike, pa je treba opozoriti na izredno previdnost, da zapis ne zaide v nasprotno smer, to je v pretirano iskanje skupne standardizirane oblike, ko to ne bi bilo upravičeno, na primer v Gosu *žiher* (v pomenu *lahko*) ni enako kot *ziher* (v pomenu *varno; zagotovo*) ...

Kot ugotavljamo že pri zapisovanju polverbalnih in neverbalnih glasov, transkribiranje izjemoma hote ali nehote zaide na spolzek teren interpretacije funkcij/pomena posameznih izrazov. Tako se v Gosu v standardiziranem zapisu pojavlja trojček izrazov *not* vs. *noter* vs. *notri*, ki pa se v govoru ne rabijo enako kot predvideva standardni jezik, tj. *notri* je lahko v vlogi standardnega *noter*, npr. *morš notri padniti*, in obratno, *noter* je v vlogi *notri,* npr. *pomijejo pa vržejo tam notri*, *not* pa lahko gledamo kot reducirano obliko enega ali drugega ali kot samostojen leksem. V obstoječem gradivu korpusa Gos se pri teh izrazih sledi interpretaciji funkcij/pomenov teh izrazov. Enako kot pri neverbalnih in polverbalnih glasovih tudi tukaj menimo, da naj ostane taka praksa čim bolj izjemna in jo je smiselno tudi za nazaj kritično pretresti od primera do primera.

6. Zaključek

V prispevku smo predpostavili, da bo morebitna nadgradnja korpusa Gos zelo verjetno potekala v obliki enotne platforme in (delno) skupnega vira z akustično bazo za razpoznavanje tekočega govora. V nadaljevanju smo se osredotočili na vprašanja zapisovanja govora, ki bi v tej skupni platformi po našem mnenju potekala na podlagi vzpostavljenega dvotirnega sistema zapisovanja (pogovorni in standardizirani zapis) v korpusu Gos. Opozorili smo na kompleksnost problema zapisovanja, ki je do neke mere vedno tudi interpretacija, ter se nato podrobneje posvetili štirim vprašanjem, ki so se nam odprla skozi uporabo in analize obstoječega Gosovega gradiva.

Že sproti smo opozorili, da je odprtih vprašanj lahko še več. Eno obsežnejših je povezano s funkcijskimi besedami in je prezapleteno, da bi ga lahko obravnavali kot del tega prispevka. Problem lahko ilustriramo s primeroma *k* in *ka*: pogovornemu *k* so v Gosu pripisane standardizirane različice *ker, ko, ki, k, kot, kjer, kar, kaj* ..., pogovornemu *ka* pa *kaj, ka, ker, da, ki, ko, kar* ... Osrednje vprašanje je, kdaj je neki reducirani obliki smiselno iskati interpretacijo v obstoječem (pisnem) standardu in kdaj jo obravnavati kot novo obliko/funkcijsko besedo. V zvezi s tem se kaže potreba po poglobljeni celostni jezikoslovni oz. jezikoslovno-diskurzni analizi rabe funkcijskih besed v govorjenem jeziku, šele potem lahko razmišljamo o prenovljenih ali dopolnjenih navodilih za standardizirani zapis te skupine besed.

V začetku prispevka smo opozorili na majhno zanimanje raziskovalcev, zlasti jezikoslovcev, za vprašanja govorjenega jezika in prepuščenost njegovega opisovanja posameznim iniciativam zunaj stroke. Kot da je to pojav, ki nam je preblizu, da bi se nam zdel neznan in zato zanimiv ter potreben analize in opisa. Toda ravno zato, ker nam je tako blizu, lahko pove veliko o človeku, več, kot se zavedamo ... če se le dovolj poglobimo vanj; tudi (ali pa celo predvsem) s pomočjo transkripcij in posnetkov v obliki govornega korpusa in baze.

7. Literatura

Dobrovoljc, H., Erjavec, T., Krek, S., Snoj, M., Verdonik, D., Vintar, Š., 2014. AKcijski načrt za jezikovno opremljenost. Dostopno na: http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/slovenski_jezik/Akcijska_nacrta/Akcijski_nacrt_za_jezikovno_opremljenost_javna_razprava.pdf. 1. julij 2014.

Verdonik, D., Kosem, I., Zwitter Vitez, A., Krek, S., Stabej, M., 2013. Compilation, transcription and usage of a reference speech corpus: The case od the Slovene corpus GOS. *Language Resources and Evaluation*, 47/4, 1031-1048.

Verdonik, D., Zwitter Vitez, A., 2011. *Slovenski govorni korpus Gos*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Žgank, A., Rotovnik, T., Verdonik, D., Kačič, Z., 2004. Baza Broadcast News za slovenski jezik (BNSI) in sistem za razpoznavanje tekočega govora. Zbornik konference IS'04 – Jezikovne tehnologije. Dostopno na: http://nl.ijs.si/isjt04/zbornik/. 1. julij 2014.

Žgank, A., Rotovnik, T., Grašič, M., Vlaj, D., Kačič, Z., 2006. Slovenska govorna in tekstovna baza parlamentarnih razprav za avtomatsko razpoznavanje govora. Zbornik konference IS'06 – Jezikovne tehnologije. Dostopno na: http://nl.ijs.si/is-ltc06/proc/22_Zgank_2of2.pdf.

Žgank, A., Zwitter Vitez, A., Verdonik, D., 2014. The Slovene BNSI Broadcast News database and reference speech corpus GOS: Towards the uniforme guidelines for future work. Zbornik konference LREC 2014. Dostopno na: http://www.lrec-conf.org/proceedings/lrec2014/index.html. 1. julij 2014.

Žibert, J., Mihelič, F., 2004. Development, evaluation and automatic segmentation of Slovenian Broadcast News Spech Database. Zbornik konference IS'04 – Jezikovne tehnologije. Dostopno na: http://nl.ijs.si/isjt04/zbornik/. 1. julij 2014.

Priloga: Predlog zapisovanja najpogostejših neverbalnih in polverbalnih glasov

Seznam je narejen na podlagi korpusa Gos v obsegu 1 mio. besed, dostopnega na www.korpus-gos.net, marca 2014. Znak # pred zapisom pomeni, da zapis ni enoznačen in je lahko identičen zapisu kake druge besede, npr. veznika, člena ipd. Če pred zapisom ni znaka #, pomeni, da se mu lahko avtomatsko pripiše enaka lema, kot je obstoječi zapis, in oblikoskladenjska oznaka za medmet.

#a
aa (zanikanje)
aaa
aam
aan
ah
aha
ahah
ahaha
ahahaha
ahja
ahjoj
ahm
ahoj
#aj
#aja
ajah
ajaj
aje
ajej
ajo
ajoj
alo
ao
aua
auva
av
#ba
bljeh
brum
bu
bvak
buf
bum
bumč
bvum
bzz
bž
ck
dammm
dh
dum
#e
eee
eem
een
eev
eh
ehe
eheh
ehehe
ej
eje
ejo
ejoj
fuf
fuj
fuu
grr
ha
haha
hahaha
hahahaha
hajaj
#he
heh
hehe
hehehe
hej
hhh
#hi
hihi
hijaj
hijo
hjoj
hjujujujuju
hm
#ho
hoho
hohoho
hohop
hojoj
hopa
hopla
hopsasa
hov
hu
huh
huhu
#i
iii
ija
ijo
ijoj
jah
jaj
jao
jea
jee
#jej
jes
johoho
johoj
joj
jojojojojojo
joo
jov
joz
juhej
juhu
juhuhu
jupi
juu
klink
maa
mahh
mee
mh
mhm
miu
mjav
mm
(pritrjevanje)
mmm
nanananananana
nee
nhn
njam
njm
nn (zanikanje)
nnn
#o
oa
oh
ohja
ohjej
oho
ohoho
ohohoho
#oj
oja
ojej
ojla
ojoj
ojojej
ojojo
ojojoj
ojojojo
ojojojoj
ojojojojoj
ola
ooa
ooo
op
opa
opala
ops
ov
ovh
paf
pavf
pff
pha
plop
pom
puf
ratatatatata
rc
rrr
ssk
sss
tada
tadadada
tadam
tarararata
taratatam
tarararan
tarararararararar
arara
taratataratat
tk
totrolodontodo
tp
tralala
tumbapa
tup
#u
ua
uf
uh
#uhu
uhuhu
ujej
umbapa
#uo
ups
upsala
vaa
vav
vov
zk
šink
šk
ššš
čk
čuf
čuči

Razvoj zbirke slovenskega emocionalnega govora iz radijskih iger - EmoLUKS

Tadej Justin[1], **France Mihelič**[1], **Janez Žibert**[2]

[1] Univerza v Ljubljani,Fakulteta za elektrotehniko, LUKS, Tržaška 25, 1000 Ljubljana
{tadej.justin, france.mihelic}@fe.uni-lj.si
[2] Univerza na Primorskem, Fakulteta za matematiko, naravoslovje in informacijske tehnologije, Glagoljaška 8, 6000 Koper
janez.zibert@upr.si

Povzetek

V tem delu predstavljamo gradnjo slovenske zbirke emocionalnega govora za namen umetnega tvorjenja govora in razpoznavanja emocionalnih stanj govorca. V prispevku se osredotočamo na opis razvite metodologije in razvoj programske opreme za označevanje paralingvistične informacije v govoru na primeru označevanja emocionalnih stanj v slovenskih radijskih igrah. Govorno zbirko in programsko opremo za množično označevanje smo v celoti zasnovali v Laboratoriju za umetno zaznavanje, sisteme in kibernetiko na Fakulteti za elektrotehniko v Ljubljani. Zbirka vsebuje govorne zvočne signale, ki so del sedemnajstih radijskih iger, s katerimi razpolagamo z licenco za akademsko uporabo. Za namen testiranja razvite aplikacije namenjene množičnemu označevanju posnetkov v tem prispevku poročamo o označeni zbirki ene govorke in enega govorca. Označevanje emocionalnih stanj je označilo pet prostovoljcev. S pomočjo razvite spletne aplikacije, ki temelji na sistemu za urejanje vsebin CMS Plone je bilo označenih 1110 posnetkov. Dodatno v prispevku poskušamo predstaviti problematiko povezano z označevanjem govornih zvočnih posnetkov na primeru množičnega označevanja emocionalnih stanj v govoru iz govornih posnetkov radijskih iger in poročamo o ujemanju oznak označevalcev.

Development of emotional Slovenian speech database based on radio drama - EmoLUKS

In this article we present the development of the Slovenian emotional speech databases developed for purposes of speech synthesis and automatic emotion recognition. The main focus in this article is about the development of methodology and software used to label the paralinguistic information from speech. The design of the database and development of the software for crowd-sourcing was produced and developed at the Laboratory of Artificial Perception, Systems and Cybernetics at Faculty of Electrical Engineering of Ljubljana. The curently annotated database consists of speech signals extracted from 17 radio dramas, with the academic licence for processing and annotating the audio signals authorized by from RTV Slovenia. For purposes of testing the developed crowd-sourcing software we focused in labelling emotional speakers states of one male and one female speaker. The emotional labels were annotated using the developed web based application with five volunteers. In this article we present the implementation of web based application for crowd-sourcing based on CMS Plone and annotating procedure which results in emotional speech database consisting of 1110 recordings. We additionally focus in the problems of annotating the speech corpora in the crowd-sourcing environment for annotating the paralinguistic informations from speech and on the example of the annotated database we report about the obtained annotations based on annotators majority vote.

1. Uvod

Dolgoletni razvoj sistemov za razpoznavanje in umetno tvorjenje govora (sintezo govora) je dodobra izpopolnil metode in principe obeh področji. Vseeno pa še vedno ostaja prostor za razvoj sistemov, ki omogočajo razpoznavanje ter tvorjenje paralingvističnih stanj govorca (Yamashita, 2013). V zadnjem desetletju je veliko pozornosti namenjene raziskavam in razvoju sistemov, ki omogočajo razpoznavanje emocionalnih stanj govorca (Ayadi et al., 2011), kakor tudi sistemov namenjenih tvorbi emocionalnega govora (Krstulovic, 2007). Dan danes so na globalnem trgu že prisotne naprave ter aplikacije, ki omogočajo interakcijo človek-stroj (ang. human-computer-interaction, HCI) preko govora. Tovrstne aplikacije so močno odvisne od jezika, kar je tudi eden od razlogov, da je opravljanje tovrstnih aplikaciji v Slovenskem jeziku mnogokrat zapostavljeno. V želji, da bi aplikacijam omogočili bolj naravno tvorbo umetnega govora ter pripomogli k raziskovanju avtomatskega razpoznavanja emocionalnih stanj, je nedvomno eden izmed prvih korakov k raziskovanju tovrstnih aplikativnih sistemov gradnja od jezika odvisne govorne podatkovne zbirke z dodatnimi paralingvistični oznakami govorca.

Do danes je bilo razvitih veliko tuje jezičnih govornih zbirk, ki skušajo zajeti tudi paralingvistična stanja govorca (Schuller et al., 2013). Tovrstna stanja se v literaturi opisujejo kot stanje govorca, katero se ne da opisati z lingvističnimi ali fonetičnimi oznakami. Paralingvistična stanja so lahko izražena v govoru kot na primer zastrupljenost, razpoloženje, zanimanje, emocionalno stanje, itd.. Načrtovanje izgradnje tovrstnih podatkovnih zbirk zahteva zahtevno interdisciplinarno sodelovanje. Eden izmed pomembnejših dejavnikov predstavlja prav opredelitev paralingvističnih oznak, kjer je nujno potreben ekspert začrtanega področja uporabe podatkovne zbirke. Tako na primer označevanje emocionalnih stanj v govoru, predstavlja težavno nalogo, saj trenutno ne razpolagamo z splošno uveljavljeno metodologijo opisovanja emocionalnih stanj. V takem primeru se velikokrat raziskovalci zatečejo k utečenim postopkom izgradnje govornih podatkovnih zbirk po zgledih v svetovni literaturi, ki opredeljuje natančne opise emocionalnih stanj v govoru in se glede na potrebe raziskovanja močno razlikujejo. V splošnem lahko potrebne opise emocionalnih stanj delimo glede na zastavljen cilj uporabe. V literaturi (Cowie and Cornelius, 2003) lahko zasledimo tipične raziskovalne potrebe, ki narekujejo smernice k izgradnji govornih emocionalnih podatkovnih zbirk in so predvsem osredotočene k raziskovalnemu cilju. Gradnjo takih podatkov-

nih zbirk največkrat usmerijo raziskovalni cilji ki strmijo k raziskovanju teoretskega ozadja emocionalnih stanj v govoru in so v večini primerov psihološke ali biološke narave. Na drugi strani so zastopani tudi cilji, ki strmijo k uporabi tovrstnih podatkovnih zbirk v aplikativne namene.

V slovenskem prostoru so do danes prisotne dve govorni zbirki emocionalnega govora, katerih opise najdemo v (Gajsek et al., 2009) in (Hozjan et al., 2002). Prva predstavlja multi-modalno zbirko spontanih emocionalnih stanj in je njena uporaba za namen sinteze govora zelo omejena. Druga predstavlja del večjezične govorne zbirke Interface, ki je dostopna pod komercialno licenco.

V tem prispevku se osredotočamo na gradnjo emocionalne govorne podatkovne zbirke za aplikativno uporabo v namen sinteze slovenskega emocionalnega govora. Predstaviti želimo dosedanje delo in probleme s katerimi se srečujemo oblikovalci tovrstnih podatkovnih zbirk. Govorna zbirka, ki jo predstavljamo v tem prispevku je izdelana preko že zajetih govornih posnetkov slovenskih radijskih iger.

Prispevek delimo na štiri poglavja, kjer v metodologiji predstavimo potrebne aplikacije za izgradnjo emocionalnih podatkovnih zbirk. Nadaljujemo z rezultati, ki predstavijo označen emocionalni govorni material ter hkrati posvetimo posebno pozornost ujemanju mnenj označevalcev. V naslednjem poglavju skušamo povzeti probleme pri gradnji tovrstne zbirke. V zaključku prestavimo nadalje delo ter komentiramo označeni del podatkovne zbirke za ciljno uporabo v sistemu za slovensko emocionalno umetno tvorjenje govora.

2. Metodologija

Dan danes je uspešnost avtomatskih sistemov, ki uporabljajo algoritme s področja umetne inteligence ter strojnega učenja, močno odvisna od velikega števila vzorcev (učna množica), ki so na razpolago za učenje modela. Uspešnost se določi s pomočjo postopkov za evalvacijo ter s pomočjo vzorcev, ki so na razpolago za testiranje (testna množica). V evalvacijskem postopku največkrat primerjamo rezultate udejanjenega sistema ter označbe vzorcev, ki so pripisane testnim vzorcem. S pridobljeno uspešnostjo lahko rečemo, da udejanjeni sistem lahko dobro ali slabo opravlja svojo nalogo z uspešnostjo tudi na naravnih vzorcih, ki niso del podatkovne zbirke na podlagi katere smo ga razvili. Dobra strategija pri izdelavi podatkovne zbirke namenjene tako učenju ter testiranju sistemov za specifično nalogo je torej ključnega pomena za udejanjanje splošnih sistemov za določeno nalogo.

V primeru pridobivanja govornih podatkovnih zbirk, namenjenih razpoznavanju in/ali tvorjenju umetnega govora, lahko opazimo, da so močno odvisne od jezika. Želja vseh razvijalcev s tega področja je pridobiti tako podatkovno zbirko, ki zajema čim več jezikovnih prvin, tako iz pisane besede kot tudi iz glasoslovja. Glavna razlika med govornimi podatkovnimi zbirkami namenjenimi umetnemu tvorjenju govora ali razpoznavanju govora, je v številu govorcev zajetih v podatkovni zbirki. V prvem primeru si v splošnem želimo razpolagati z obsežno zbirko enega govorca, ki vsebuje čim večje število različnih posnetkov. V drugem primeru pa si želimo razpolagati z zbirko čim večjega števila različnih govorcev. Tovrstne podatkovne zbirke ponavadi vsebujejo posnetke enakih v naprej predvidenih stavkov prebranih s strani več govorcev. S takimi strategijami v prvem primeru pridobimo dovolj raznolik in čim boljši približek govorjene besede (jezika) posameznega govorca. V drugem primeru, pa si želimo razviti čim bolj robusten model, ki omogoča dobro razpoznavanje čim večjega števila uporabnikov.

Razvijalci emocionalnih govornih podatkovnih zbirk, ki so namenjene za aplikativno rabo v avtomatskih sistemih za umetno tvorjenje govora ali razpoznavanje emocionalnih stanj govorca, se velikokrat poslužujejo dveh strategij, ki omogočata zajem zbirke. Prvi predstavlja snemanje govorne zbirke s pomočjo profesionalnih bralcev, ki so zmožni posnemati emocionalna stanja. Take zbirke so posnete z pomočjo vnaprej pripravljenih povedi, ki so izbrane iz obsežnejših zbirk besedil in v njihovi celoti skušajo zadostiti fonetični porazdelitvi osnovnih enot posameznega jezika. V drugem primeru pa razvoj zbirke zajema pridobivanje že posnetkih govornih segmentov, ki jih je potrebno točno in natančno prepisati ter v primeru večjega števila govorcev na posnetku dodatno časovno določiti identiteto govorca v posameznem posnetku. V obeh primerih je potrebno govorne posnetke označiti z vnaprej predvidenimi emocionalnimi oznakami. V zadnjem času se za to nalogo velikokrat najame označevalce, katerih večinsko mnenje določa končno oznako posameznega posnetka. V primeru podatkovne zbirke EmoLUKS predstavljajo označeni signali igrana emocionalna stanja govorcev, saj so radijske igre posnete z poklicnimi igralci.

2.1. Transkripcija in segmentacija govornih posnetkov

S pomočjo RTV Slovenija smo pridobili radijske igre, ki so bile v večini posnete v profesionalnem studiu radia Slovenija. Vsako radijsko igro smo transkribirali ter razčlenili glede na identiteto govorca. V poteku segmentacije in transkribiranja smo želeli označiti predvsem čisti govor, zato smo vzporedno označevali tudi nejezikovne prvine, ki so večkrat del radijskih iger. To se v večini glasba v ozadju, različni šumi ter raznovrstni dodatni zvočni efekti. Poleg tega nismo pozabili tudi ostale nejezikovne prvine govorca, kot so vdih, cmokanje, stokanje, jok in smeh.

Za potrebe prepisov in razčlenitve glede na govorca smo uporabili programa Transcriber (Barras et al., 2001). Orodje omogoča hitro in učinkovito razčlenjevanje govornih signalov glede na govorca, njihovo transkripcijo ter označevanje ne jezikovnih elementov v posnetku. Posnetke smo razčlenili tudi glede na zaključene stavčne enote. S takim pristopom smo pridobili nabor posnetkov, ki niso predolgi in hkrati nudijo dovolj konteksta za označevanje paralingvistične informacije v govoru.

Z orodjem Transciber smo označili 17 posnetkov radijskih iger v približnem skupnem časovnem obsegu 12 ur in 50 minut. Tabela 1 na strani 3 prikazuje količino transkribiranega in označenega materiala.

2.2. Definicija emocionalnih stanj

Vsakršno raziskovalno delo, ki posega na področje čustev, potrebuje najprej definicijo, kaj čustvo je oziroma, kaj

št.	Naslov radijske igre	Trajanje
1	Penzion Evropa	0:48:03,56
2	Angleško poletje	0:57:55,69
3	V Sieni nekega deževna dne	0:42:32,59
4	Aut Caesar	0:33:22,25
5	Štefka	0:36:45,69
6	Podzemne Jame	0:46:17,56
7	Na glavi svet	0:58:29,32
8	Nas novi najboljši prijatelj	0:26:51,25
9	Dediščina	0:54:36,62
10	Potovalci	0:49:50,27
11	Nič brez Deteljnika	0:48:00,00
12	Sokratov zagovor	1:09:51,34
13	Nedotakljivi – Četrti žebelj	0:38:44,35
14	Nedotakljivi – Moj ded Jorga Mirga	0:37:00,00
15	Nedotakljivi – Moj oče Ujaš Mirga	0:40:04,45
16	Nedotakljivi – Jaz, Lutvi Belmondo aus Shangkai Gav	0:35:09,83
17	Hipopituitarizem ali namišljeni bolnik	0:46:50,35
skupaj		**12:50:25,12**

Tabela 1: Pregled trajanj celotnih radijskih iger

s širokega področja analize čustev pri človeku bo središče preučevanja. Razlikovanja v teoretskih predpostavkah na katerih temelji teorija o čustvih (Cornelius, 1996), pričajo o razlikovanju tolmačenja čustev. V literaturi se pojavljajo štirje različni pogledi na čustveno stanje (Cornelius, 2000). Imenujemo jih Darvinistični pogled, pogled po Jamesu, kognitivistični pogled in socialno-konstruktivistični pogled. Preko različnih pogledov na čustvena stanja se posledično uporabljajo tudi različni modeli, ki opisujejo relacije med različnimi čustvenimi kategorijami. Osnovna predpostavka, na kateri temeljijo razmejitve med čustvenimi kategorijami pri vseh modelih, je da so razlike med opaženimi čustvenimi doživljaji znotraj ene kategorije manjše od razlik med tistimi iz različnih kategorij. Pregled modelov čustev je predstavljen v (Cornelius, 1996), kjer avtor predstavlja modele znotraj štirih glavnih skupin in jih imenuje prostorski, diskretni, pomenski in komponenti modeli.

V tem prispevku se avtorji osredotočamo na diskretizacijo čustvenih stanj po Darwinovem pogledu na čustvena stanja. Tako se osredotočamo na predpostavko, da obstaja nekaj osnovnih čustev, iz katere so se razvili osnovni diskretni modeli čustvenih kategorij. Tak pristop je tudi eden izmed najpopularnejših predstavitev prostora čustvenih stanj. Na tak način smo diskretizirali osnovna čustvena stanja v naslednje kategorije, ki smo jih uporabili za označevanje posnetkov radijskih iger: žalost, veselje, gnus, jeza, strah, presenečenje in nevtralno.

Označevanje emocionalnih stanj v govoru poteka s pomočjo ekspertnega znanja. Govornim posnetkom lahko pripiše oznako ekspert za dano področje. V zadnjem času se večkrat uporablja nabor označevalcev, ki podajo svoje mnenje o posameznem posnetku. S takim naborom mnenj lahko bolj posplošeno določimo oznako posnetku. Ker je označevanje govornih posnetkov velikokrat dolgotrajen proces, se za označevanje podatkovnih zbirk večkrat uporablja spletne aplikacije, ki omogočajo hkratno označevanje večjega števila označevalcev ter hkrati ponujajo označevalcem svobodno izbiro časovnega okvira označevanja. V literaturi tovrstni pristop zasledimo pod pojmom množično izvajanje (ang. croud-sourcing) (Howe, 2006).

2.2.1. Razvoj in opis aplikacije za označevanje zvočnih in video posnetkov

Po pregledu literature ter ogledom dostopnih spletnih aplikacij, ki nudijo označevanje govornih posnetkov, smo se odločili, da nobena v taki meri ne izpolnjuje pogojem, ki bi morali biti zadoščeni, da lahko enostavno ponudimo našim prostovoljnim označevalcem kvalitetno in hitro označevanje. Zato smo se odločili izdelati spletno aplikacijo namenjeno označevanju zvočnih ali video posnetkov. K taki odločitvi nas je napeljalo tudi dejstvo, da lahko razpolagamo in obdelujemo tovrstne podatke samo za akademske potrebe. Zato bojazen, da ob prenosu na spletno mesto večjih razsežnosti ter posledični kraji intelektualne lastnine, ki nam je bila zaupana v varstvo s strani Radia Slovenija zadostimo le v primeru, če omogočimo gostovanje posnetkov tovrstne aplikacije na lastnih spletnih strežnih. Z uporabo sodobnih tehnologij ter nenehnim varovanjem in nadzorovanjem strežniškega sistema se lahko kraji takih podatkov izognemo, vendar le v primeru, če razpolagamo z popolnim nadzorom nad aplikacijo in strežnikom.

Spletno aplikacijo smo razvili s pomočjo odprto kodnih prosto dostopnih programov. Spletna aplikacijo je bila razvita kot dodatek k sistemu za urejanje vsebin (ang. Content Management System, CMS) Plone verzije 4.3.3 [1]. Odprto kodni sistem CMS Plone je razvit na podlagi programskega ogrodja za urejanje vsebin Zope [2]. Izbira takega osnovnega sistema za razvoj aplikacije sloni na dejstvih, da je sistem Plone eden izmed spletnih CMS, ki se ponaša z eno boljših sledi o zapisih varnostnih popravkov [3] ter je zaradi tega uvrščen v skupino najbolj varnih CMS. Poleg tega ima že vgrajen način za delo z delovnimi tokovi (ang. workflows), ki so nujno potrebni za enostavno urejanje nad pravicami za ogledovanje, urejanje in ustvarjanje spletnih vsebin. Hkrati ponuja razvoj dodatkov, ki jih lahko razvijalec implementira in namesti v že obstoječi sistem.

Spletno aplikacija sestoji iz uredniških in uporabniških strani. Tako uredniki kot tudi uporabniki (ocenjevalci) se morajo v sistem prijaviti z uporabniškim imenom in geslom. Ker je označevanje zvočnih ali video posnetkov lahko dolgotrajen proces, smo s takim pristopom zagotovili označevalcem možnost označevanja v daljšem časovnem obdobju. Hkrati smo onemogočili naključnim obiskovalcem dostop do občutljivih posnetkov. Spletna aplikacija namenjena označevanju zvočnih ali video posnetkov je dostopna na http://emo.luks.fe.uni-lj.si.

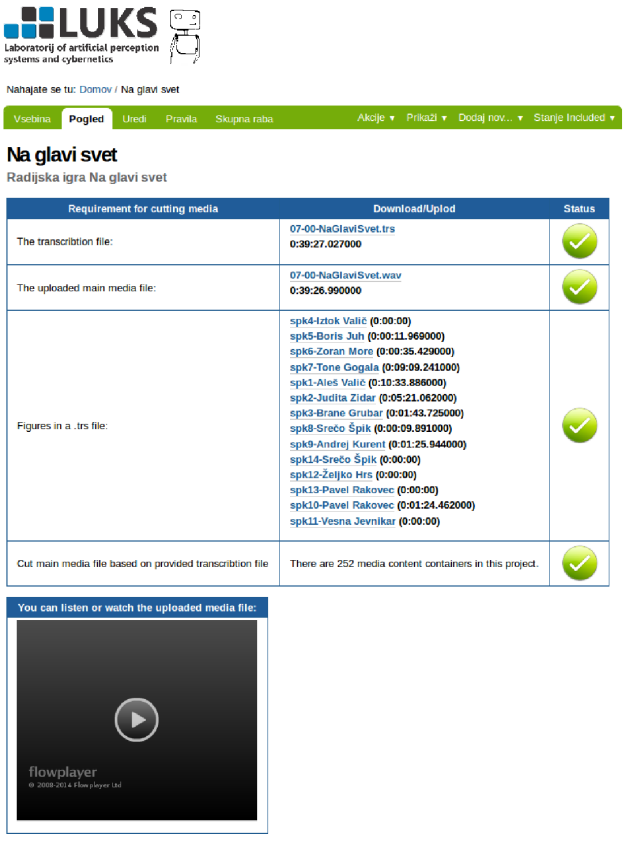
2.2.2. Uredniške strani

Poleg preprostega označevanje paralingvistične informacije v zvočnih ali video posnetkih, ki jo nudi spletna aplikacija smo pripravili tudi uredniški vmesnik, ki omogoča enostavno in hitro izdelavo projekta označeva-

[1] http://plone.org
[2] http://zope.org
[3] http://cve.mitre.org/

nja. Uredniku je omogočen enostaven prenos obsežnejše zvočne ali video datoteko s pripadajočo datoteko v tekstovni obliki, ki vsebuje potrebne zapise o segmentaciji in transkribciji. Trenutno podprt format datoteke za segmentacijo je format Transcriber XML. Spletna aplikacija avtomatsko razreže posnetek na manjše zaključene posnetke, ki so časovno označeni v datoteki za segmentiranje. Določi jim tudi identiteto govorca ter jasno razdeli posnetke s čistim govorom ter posnetki, ki vsebujejo poleg govora tudi druge nejezikovne prvine. Aplikacija poskrbi tudi za pravilen format prikaza zvočnih ali video posnetkov v različnih spletnih brskalnikih. Slika 1, prikazuje uspešen uvoz potrebnih podatkov na spletni strežnik.

Uredniku spletne aplikacije je po uspešnem uvozu podatkov omogočena enostavna izdelava projekta označevanja. Urednik z vnosom zahtevanih parametrov ustvari nalogo namenjeno označevanju. V tem delu ima možnost vključiti katere večje enote posnetkov bodo na voljo za označevanje. V našem primeru so to radijske igre. Uredniku je na tem mestu omogočeno, da lahko na enostaven način vključi v nalogo označevanja le del vseh dostopnih sklopov podatkov na strežniku. Poleg tega lahko urednik enostavno izbira iz nabora vseh identitet govorcev le tiste, za katere meni, da je označevanje smiselno. V tem primeru bodo v nalogo označevanja vključeni le posnetki, ki vključujejo govor vključenih identitet govorcev. K začetni inicializaciji in selekciji govorcev v postopku ustvarjanja naloge označevanja sodi tudi podroben opis naloge označevanja ter opis samega procesa označevanja. Uredniku je omogočeno preko pisane besede, slik in povezav v splet jasno predstaviti označevalcu cilje označevanja in hkrati pregledno opisati in definirati katere paralingvistične oznake so na voljo za označevanje posnetkov. V posebno polje urednik vnese skrajšana imena z nabora opredeljenih paralingvističnih oznak. Ta imena so kasneje uporabnikom v procesu označevanja prikazana kot možna izbira za označbo posnetka. Ker aplikacija zahteva tudi uvoz transkripciji lahko urednik omogoči izpis prepisa zvočnega posnetka. Ker so v večini paralingvistični dejavniki v govoru zaznani preko širšega konteksta aplikacija omogoča poleg prikaza transkribcije samega posnetka tudi izpis širšega nabora transkribciji pred in po posnetkom, ki ga označuje označevalec. Tak pristop označevalcu nudi vpogled v predhodno ter kasnejše dogajanje, ki omogoča označevalcu umestitev posnetka v širši kontekst. V primeru označevanja radijskih iger se to izkaže za koristno, saj so velikokrat prisotni dialogi med dvema govorcema in s tako ponazoritvijo dialoga označevalcu pojasnimo kontekstno dogajanje.



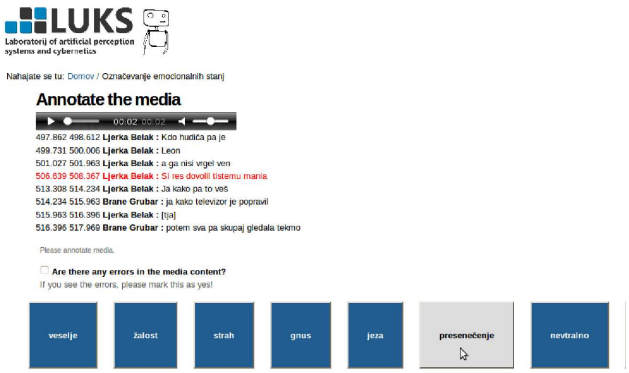
Slika 1: Uredinške strani aplikacije za označevanje, primer uspešnega vnosa podatkov.

Pri dolgotrajnem procesu transkribiranja in segmentiranja zvočnih ali video posnetkov se velikokrat pojavijo napake. Urednik spletne aplikacije ima možnost omogočiti sistem poročanja o napakah. Ta sistem označevalcem omogoča enostavno označbo, da posnetek vsebuje napako. Taki posnetki so uredniku prikazani ločeno, ki jih lahko enostavno popravi s pomočjo urejanja posnetka kar preko spleta.

V primeru razvoja paralingvističnih govornih podatkovnih zbirk iz že vnaprej posnetega govornega materiala so razvijalci primorani material, ki ga hočejo vključiti podatkovno zbirko v celoti pregledati in žal tudi nekaj material, ki ni smiseln ali pa je posledica nenatančne sementacije v postopku transkribiranja in segmentiranja zavreči. Aplikacija zato urednikom ponuja pred objavo procesa označevanja izvedbo testnega protokola označevanja. Uredniku je ponujena možnost identičnega označevanja, ki je kasneje na voljo posameznemu označevalcu. Med potekom testa ima urednik možnost enostavne izključitve posnetka iz procesa označevanja.

2.2.3. Uporabniške strani

Po prijavi v spletni sistem je označevalec najprej seznanjen z nalogo in opisom protokola označevanja zvočnih ali video posnetkov. Z sprejetjem pogojev označevanja lahko označevalec prične z delom. Na zaslonu sem mu avtomatsko predvaja govorni ali video posnetek iz nabora posnetkov, ki so del označevalnega procesa. Vrstni red posnetkov, ki jih označevalec označuje je naključen. Označevalec s klikom na gumb pod posnetkom izbere med možnimi izbirami oznak. Po izbiri oznake se mu na zaslon prikaže naslednji posnetek namenjen označevanju in se avtomatsko predvaja. V kolikor se označevalec zmoti ima vedno možnost popravka zadnjih pet odločitev. Slika 2 prikazuje označevanje posnetka za osnovna emocionalna stanja. V kolikor urednik omogoči dodatne opcije, se pod predvajalnikom posnetka izpiše tudi širši kontekst prepisov, kateremu sledi tudi možnost označbe napake v posnetku ali v transkribciji. Sledijo možne izbire posameznih vnaprej predvidenih emocionalnih stanj, ki so v našem primeru na voljo za označevanje.



Slika 2: Uporabniška stran v postopku označevanja emocionalnih stanj.

Označ.	Povprečen čas odločitve	Št.popravkov odločitev	Skupen čas odločitev
01m	26,04	14 (1,26 %)	08:01:44,08
02m	10,12	6 (0,54 %)	03:07:16,43
03m	15,10	1 (0,09 %)	04:39:27,65
01f	30,51	2 (0,18 %)	09:24:07,88
02f	31,68	10 (0,90 %)	09:46:07,88
Skupaj	**22,64**	**33 (0,59 %)**	**34:58:43,93**

Tabela 2: Označevanje 1110 posnetkov s povprečnim trajanjem 4,13 sekunde iz sklopa 17 radijskih iger petih označevalcev. Prvi stolpec označuje identiteto označevalca ter spol.

Aplikacija ima vgrajen tudi sistem za merjenje časa, ki ga označevalec porabi za odločitev. Ker ima označevalec vedno možnost večkratnega poslušanja enega posnetka lahko z analizo takih podatkov hitro ugotovimo najbolj problematične odločitve. Hkrati nudi kontrolo nad najmanjšim časovnim obsegom, ki ga mora poslušalec nameniti za izvedbo odločitve. Porabljen čas mora biti vedno večji, kot pa čas posnetka predvajanja.

Ko označevalec označi vse predvidene posnetke se proces označevanja zaključi. Takrat je označevalcu ponujen vpogled v porabo časa, ki ga je namenil za označevanje in hkrati vpogled v kratek pregled števila označenih posnetkov.

2.3. Postopek označevanja emocionalnih stanj pri zbirki Emo LUKS

V tem prispevku predstavljamo trenutno označeno delo, ki obsega govorca ženskega in moškega spola. V nabor označevanja smo vključili vse radijske igre. Ocenjevalcem je bilo omogočeno poročanje napak ter tudi izpis širšega konteksta transkribcije posnetka. V označevalnem procesu je bilo vključenih 1110 posnetkov v skupnem časovnem obsegu 1 ure 16 minut in 24 sekund. Povprečni čas trajanja posnetka, ki je vključen v proces označevanja je 4,12 sekunde. V procesu označevanja je sodelovalo pet označevalcev. Trije moški in dve ženski. Vsi označevalci so uporabljali slušalke. Vsak od označevalcev je označil vse posnetke. V tabeli 2 so prikazani povprečne vrednosti časa potrebnega za izvedbo odločitev, število popravkov odločitev označevalca ter skupen efektivni čas označevanja vseh 1110 posnetkov.

Vsak označevalec je lahko izbiral med osnovnimi kategorijami emocionalnih stanj; žalost, veselje, gnus, jeza, strah, presenečenje in nevtralno. Poleg osnovnih stanj govorca je bila dodana tudi kategorija ”nič-od-tega”, ki označevalcu omogoča označbo emocionalnega stanja govorca, ki ni del predpisane sistemizacije kategorij osnovnih emocionalnih stanj v govoru.

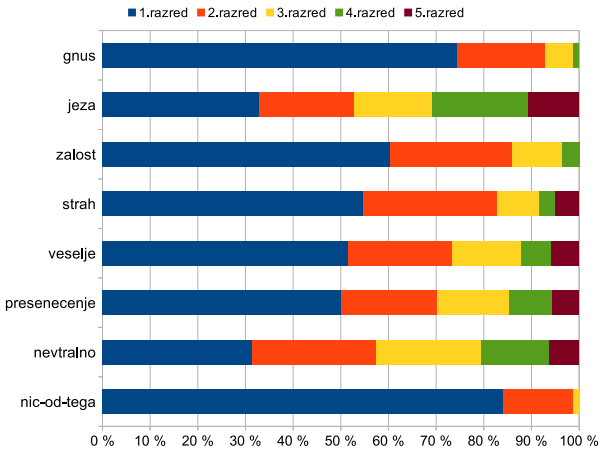
3. Rezultati

Spletna aplikacija omogoča enostaven izvoz vseh označenih podatkov v tekstovno obliko formata ”.csv” (ang. coma separated value). Tak format podpira večina programskih orodji za statistično obdelavo in analizo podatkov.

Ujemanja mnenj označevalcev prikazuje slika 3. Na sliki so prikazani deleži odločitev v petih razredih. Vsak izmed razredov predstavlja razmerje števila mnenj označevalcev o posameznem posnetku v posamezni emocionalni kategoriji proti številu vseh mnenj za posamezno emocionalno kategorijo. Na ta način je v prvem razredu zastopan delež posnetkov za katere je en označevalec podal mnenje o posamezni kategoriji. V drugem razredu so zastopani deleži posnetkov, pri katerih sta se dva označevalca odločila za enako mnenje. Analogija o definiciji posameznih razredov se lahko enostavno razvleče do petega razreda.

Iz slike 3 lahko razberemo, da je ujemanje vseh petih ocenjevalcev o posameznem posnetku (5. razred) zastopano pri označbah jeza, strah, veselje, presenečenje in nevtralno. Delež ujemanja vseh petih ocenjevalcev pri navedenih kategorijah označb se giblje od 4,9% do 10,5%. Ker so deleži popolnega ujemanje premajhni in bi z takim pristopom določevanja oznak posnetkov pridobili le malo število posnetkov, nekatere kategorije pa bi bili primorani celo zavreči se končna oznaka posameznega posnetka določi s pomočjo večinskega mnenja označevalcev (ang. majority voting).

Tabela 3 prikazuje pridobljen material s pomočjo večinske odločitve označevalcev na posameznem posnetku. Tre-



Slika 3: Slika ujemanje mnenj o posameznih posnetkih.

Govorec	Št. posnet.	Trajanje	Deleži oznak čustvenih stanj [%]								
			nev.	jeza	ves.	pres.	žal.	strah	gnus	ned.	nič
01m_av	762	01:01:28,60	36,48	14,44	8,53	11,02	4,86	5,38	1,18	16,67	1,44
01f_lj	348	14:55,55	11,21	28,45	11,78	14,94	2,30	8,05	4,02	17,82	1,44
skupaj	1110	01:16:24,15	28,56	18,83	9,55	12,25	4,05	6,22	2,07	17,03	1,44

Tabela 3: Pregled deležev označenih emocionalnih posnetkov s pomočjo določitve večinskega strinjanja označevalcev.

nutno razpolagamo z oznakami petih označevalcev preko vsega govornega materiala iz 17 radijskih iger, identitete govorca *av* in govorke *lb*.

4. Diskusija

Označevanje emocionalnega stanja govorca se velikokrat izkaže za težavno nalogo, saj ni splošno sprejete definicije kategorij emocionalnih stanj. To dejstvo potrjujejo tudi rezultati, saj se menja označevalcev o posnetkih velikokrat razlikujejo. Z podrobnim pregledom slike 3 na strani 5, lahko ugotovimo, da do popolnega konsenza med označevalci prihaja le v redkih primerih. Vseeno lahko potrdimo da je v petih od sedmih klasifikacij med označenimi oznakami emocionalnih stanj v deležu med približno 5% in 10% procentov prisoten popolni konsenz med mnenji označevalcev. Slednje nakazuje na dejstvo, da so v slovenskih radijskih igrah jasno izražena emocionalna stanja igralcev in da je izbira radijskih iger smiselna za gradnjo zbirke slovenskega emocionalnega govora.

Večinsko mnenje označevalcev prikazuje tabela 3 na strani 6.Iz podatka o deležu nedoločenih oznak nad posnetki opazimo, da je 17% izmed vseh posnetkov, takih katerim ne moremo s pomočjo večinskega odločanja določiti emocionalnega stanja. Izmed vseh nedoločenih posnetkov je 91% takih, katerim sta dva označevalca pripisala eno emocionalno stanje, druga dva pa drugo emocionalno stanje in le 9% takih, katerim je vseh pet označevalcev pripisalo različno emocionalno stanje. Na tem mestu se zdi smiselno za posnetke z nedoločenim stanjem postopek označevanja ponoviti in s tem preveriti konsistentnost označevalcev ter opazovati ali posnetki resnično vsebujejo večdimenzionalna oziroma prepletajoče se emocionalna stanja.

Zasnovana spletna aplikacija se je se je izkazala za uspešno izbiro, ki omogoča hiter in učinkovit način označevanja paralingvističnih stanj v govornih ali video posnetkih tako za razvijalce podatkovne zbirke, kot tudi za označevalce. Vseeno lahko iz tabele 2 na strani 5, opazimo veliko časovno odstopanje potrebno za označevanje med ženskimi in moškimi označevalci. Razloge za to lahko iščemo v dveh dejavnikih. Prvi se nanaša na natančno označevanje z večkratnim poslušanjem govornih posnetkov ter drugi na šibko internetno povezavo, ki lahko vpliva na hitrost prenosa kratkih govornih posnetkov.

5. Zaključek

V prispevku smo predstavili spletno aplikacijo za množično označevanje paralingvistične informacije v govornih ali video posnetkih na primeru označevanja emocionalnih stanj v govoru iz slovenskih radijskih iger za namenom izgradnje govorne zbirke slovenskega emocionalnega govora - EmoLUKS.

Čeprav podatkovna zbirka vsebuje nekoliko manj govornega materiala za posamezno emocionalno stanje, kot smo to pričakovali se vseeno nadejamo, da lahko z uporabo sodobnih pristopov za tvorjenje umetnega govora s pomočjo Prikritih Markovih Model in priročnih postopkov adaptacije akustičnih modelov, ne glede na manjšo količino materiala, ki je natančno označena pripomoremo k večji naravnosti umetnega govora.

6. Literatura

Moataz El Ayadi, Mohamed S. Kamel, in Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572 – 587.

Claude Barras, Edouard Geoffrois, Zhibiao Wu, in Mark Liberman. 2001. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1–2):5 – 22.

Randolph R Cornelius. 1996. *The science of emotion: Research and tradition in the psychology of emotions.* Prentice-Hall, Inc.

Randolph R Cornelius. 2000. Theoretical approaches to emotion. V: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

Roddy Cowie in Randolph R. Cornelius. 2003. Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1–2):5 – 32.

Rok Gajsek, Vitomir Struc, France Mihelic, Anja Podlesek, Luka Komidar, Gregor Socan, in Bostjan Bajec. 2009. Multi-modal emotional database: Avid. *Informatica (Slovenia)*, 33(1):101–106.

Jeff Howe. 2006. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4.

Vladimir Hozjan, Zdravko Kacic, Asuncion Moreno, Antonio Bonafonte, in Albino Nogueiras. 2002. Interface databases: Design and collection of a multilingual emotional speech database. V: *LREC*.

Sacha Krstulovic. 2007. A review of state-of-the-art speech modelling methods for the parameterisation of expressive synthetic speech. Tehnično poročilo, DFKI Deutsches Forschungszentrum für Künstliche Intelligenz.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, in Shrikanth Narayanan. 2013. Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4 – 39.

Yoichi Yamashita. 2013. A review of paralinguistic information processing for natural speech communication. *Acoustical Science and Technology*, 34(2):73–79.

Prvi leksikalni podatki o slovenskem znakovnem jeziku iz korpusa Signor

Špela Vintar, Boštjan Jerko, Marjetka Kulovec

Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, SI-1000 Ljubljana
spela.vintar@ff.uni-lj.si, {marjetka.kulovec, bostjan.jerko}@guest.arnes.si

Povzetek

O leksikalnih in slovničnih lastnostih slovenskega znakovnega jezika (SZJ) je bilo do nedavnega mogoče pisati zgolj na podlagi domnev in posamičnih opažanj. Prispevek predstavlja rezultate leksikalne analize, ki je bila izvedena s pomočjo korpusa Signor, prvega reprezentativnega korpusa SZJ. Po predstavitvi osnovnih korpusnih statistik se v prispevku posvetimo izbranim leksikalnosemantičnim elementom SZJ, med drugim tudi vlogi mašil in gest. Opažene pogostosti primerjamo s podatki, pridobljenimi v sorodnih raziskavah v tujini, predvsem za britanski BSL in avstralski Auslan. V zaključku razpravljamo o nadaljnjih raziskavah in možnostih uporabe korpusa Signor za posodobitev jezikovnih priročnikov in razvoj jezikovnih tehnologij.

First Lexical Analysis of the Slovene Sign Language based on the Signor Corpus

The lexical and grammatical properties of Slovene Sign Language (SZJ) have so far only been described on the basis of isolated observations or presumptions. This paper presents the results of a lexical analysis performed on the Signor corpus, the first representative corpus of SZJ. After presenting general corpus statistics we discuss selected lexical and semantic properties of SZJ, for example the role of fillers and gestures. The figures obtained are compared to related works, in particular corpus-based studies performed for BSL and Auslan. The paper concludes by outlining our plans for future research and the ways in which our corpus could help improve basic reference works for SZJ as well as serve as a basis of new technologies.

.

1. Uvod

Slovenski znakovni jezik (SZJ) je jezik gluhe skupnosti v Sloveniji in ga po aktualnih ocenah uporablja od 800 do 1600 oseb. Od leta 2002 je priznan kot eden od uradnih jezikov v Sloveniji, kar uporabnikom daje pravico do sporazumevanja v SZJ v vseh javnih in zasebnih situacijah. Ta pravica se običajno udejanja prek tolmačev SZJ. Gluhi pa se večkrat počutijo zapostavljene, ker je tolmačenje drago in državni sistem vaučerjev mnogim ne zadošča, tolmači pa tudi niso vedno na razpolago. To velja še posebej za izobraževanje.

Priročniki za učenje SZJ so zasnovani brez prave jezikoslovne osnove, saj je bilo doslej tovrstnih raziskav SZJ malo. Gluhi skupnosti sta na voljo dva multimedijska slovarja, vendar noben od njiju ne temelji na korpusnih podatkih o SZJ in oba težita k strogo normativni obravnavi kretenj. Prvi učbenik za učenje SZJ je bil objavljen leta 2010 in sicer zapolnil dotedanjo vrzel v izobraževanju SZJ, vendar se učbenik močno naslanja na slovenščino in zanemarja ne le lastnosti SZJ, ampak tudi edinstvene značilnosti znakovnih jezikov na splošno.

V dosedanjih znanstvenih objavah o SZJ se avtorji večinoma ukvarjajo s sociološkim vidikom ali temami (specialne) pedagogike (npr. Kuplenik, 1999; Globačnik, 2007). Bolj jezikoslovno usmerjeni članki obravnavajo temo standardizacije (Bauman 2007), primerjavo SZJ z govorjeno/pisno slovenščino (Globačnik 2001, Žele 2007) ali opisujejo izbrane in splošne lastnosti morfologije in fonologije SZJ (Žele in Bauman 2011). Glede na pomanjkanje znanstvene osnove se pojavljajo vprašanja o kakovosti poučevanja SZJ, pa tudi priprave in certificiranja za tolmače.

Predstavljena raziskava je pomemben korak v smeri izboljševanja stanja. S projektom, ki ga je financirala Javna agencija za raziskovanje RS v obdobju 2011-2014, je bil zgrajen prvi korpus SZJ (korpus Signor, spletna stran projekta http://lojze.lugos.si/signor/index.html). Zbrali smo posnetke prek 80 uporabnikov SZJ. Podatki so uravnoteženi po regijah, starosti in spolu. Korpus je ročno tokeniziran in lematiziran, nadaljujejo pa se označevanja kompleksnejših ravni. Korpus Signor predstavlja vir za opisne raziskave SZJ v obliki, kot se ta uporablja v realnih komunikacijskih situacijah gluhe skupnosti.

Namen prispevka je opisati nekatere leksikalne lastnosti SZJ, ki se izkazujejo na podlagi osnovne ravni označevanja, se pravi na ravni pripisovanja pomenskih oznak oz. lematizacije. Za ostale raziskave, predvsem skladenjske strukture in drugih slovničnih lastnosti, bo treba počakati na nadgradnjo označevalnih ravni.

2. Korpus SIGNOR

Gradnja korpusa se je pričela leta 2011 z namenom zagotovitve reprezentativnega in uravnoteženega korpusa izvirnih primerov besedil v SZJ (prim. Vintar in dr., 2012). Pred začetkom projekta smo pregledali podobne raziskave o znakovnih jezikih po svetu: ameriškega znakovnega jezika ASL (Lu in Huenerfauth, 2011), avstralskega Auslan (Johnston in dr., 2006), avstrijskega ÖGS (Krammer in dr., 2001, Dotter, 2011) in italijanskega LIS (Prinetto in dr., 2011), vendar smo se na koncu najbolj naslonili na projekt korpusa nemškega znakovnega jezika DGS. Kljub temu, da je slovenski projekt po trajanju in sredstvih skromen v primerjavi s podobnimi, smo uporabili podobno metodologijo za izbiro informantov in strukturo snemanja (Nishio in dr., 2010), pa tudi za strategijo segmentiranja (Hanke in dr., 2012) in označevanja lem ter kompleksnih struktur (Konrad in dr., 2012).

2.1. Gradnja korpusa

Korpus je reprezentativen glede na ocenjeno velikost gluhe skupnosti v Sloveniji. Zbrali smo posnetke 80 informantov, kar predstavlja 5 do 10 % celotne skupnosti uporabnikov SZJ. Informanti prihajajo iz vseh slovenskih

regij, vzorec pa je tudi dobro uravnotežen po spolu (37 žensk in 43 moških) in starosti (leta rojstva so enakomerno razporejena od 1932 do 1996). Vsakega informanta smo prosili še za nekaj osebnih podatkov, ki so shranjeni ločeno od posnetkov: kdaj se je pojavila gluhota in njena stopnja, primarna roka, stopnja izobrazbe, mesto in regija rojstva, mesto in regija izobraževalne ustanove in uporaba slušnega aparata (prim. Vintar in dr., 2012).

Glede na to, da so uporabniki SZJ znanje jezika pridobili na različne načine in v različnih starostih, je kompetenco SZJ težko oceniti. Podobno kot v mnogih drugih družbah se znakovni jezik v Sloveniji poučuje šele v zadnjem času. Tako se starejša generacija gluhih ni učila znakovnega jezika v šoli in so bili jezikovno zanemarjani, ali pa so jih učili govoriti in odgledovati. Tudi pri mlajši generaciji so razlike v znanju SZJ velike in so odvisne od mesta šolanja, saj je v Sloveniji še vedno edina šola, kjer sistematično poučujejo SZJ, Zavod za gluho in naglušno mladino Ljubljana. Seveda je pomemben podatek tudi stopnja gluhote.

Ker smo korpus Signor gradili z namenom jezikoslovnega opisovanja SZJ, smo vprašanje kompetence rešili pragmatično. Zavzeli smo stališče, da je uporabnik SZJ tisti, ki pogosto uporablja ta jezik za primarno komunikacijo z drugimi. Podatke, ki bi lahko vplivali na kompetenco in rabo, smo shranili kot meta podatke. Tako liberalen način se je izkazal za pretežno uspešnega, res pa je, da smo pri analizi posnetkov opazili enega informanta, ki je v večji meri uporabljal govor z minimalno uporabo SZJ.

Dogovarjanje za snemanja, komunikacijo z informanti in na koncu intervju ter snemanje so izvajali izključno gluhi študentje. Nekatera snemanja so potekala v društvih in nekatera na domovih informantov. Poleg tega smo v korpus vključili posnetke dijakov na Zavodu za gluhe in naglušne Ljubljana. Pred snemanjem mladoletnih oseb smo pridobili pisno privolitev njihovih staršev.

Snemanje je potekalo v treh delih. V prvem delu je informant v znakovnem jeziku pripovedoval o svojem življenju in družini. Namen tega dela je bila tudi vzpostavitev neformalnega dialoga med spraševalcem in informantom, ki slednjemu omogoči, da se sprosti in navadi na kamero. V drugem delu je informant pred snemanjem pogledal posnetek o splošni temi (npr. politika, telo, potovanje itd.). Zadnji del snemanja je bil namenjen zbiranju strokovnega besedišča in je lahko potekal kot pogovor med spraševalcem in informantom o priljubljeni temi slednjega (hobi, šport s katerim se ukvarja) ali pa je bil predvajan posnetek z bolj specializirano tematiko, o kateri je nato tekel pogovor. Posneti pogovori s posameznimi informanti so tako trajali od 10 do 20 minut.

2.2. Obdelava korpusa in označevanje

Vsi posnetki so pretvorjeni v enoličen format (.mov) in shranjeni na projektnem strežniku. Za označevanje korpusa smo uporabili orodje iLex (Hanke in Stolz, 2008), ki predstavlja prilagodljivo večuporabniško okolje za označevanje in shranjuje vse kretnje, lekseme in druge ravni oznak v bazo. Tako smo dosegli enolično uporabo oznak pri vseh označevalcih.

Označevalna shema v sedanji različici korpusa ima več ravni:

Tokenizacija. Posnetek dialoga v znakovnem jeziku je segmentiran v posamezne kretnje in ločen s časovnimi kodami. Pri segmentaciji smo se odločili za natančnejšo različico, ki zahteva več dela, a je za nadaljno analizo bolj natančna: prehodov nismo obravnavali kot dele kretenj, tako da je med dvema označenima kretnjama večinoma časovni presledek, ki predstavlja prehod.

Lematizacija. Označevanje posameznih kretenj z leksikalnimi oznakami ali glosi ustreza lematizaciji - vsaki kretnji dodamo enolično pomensko oznako. Označevalno okolje iLex uporablja leksikalno bazo, ki vsebuje vse poznane kretnje in njihove različice. Za osnovo leksikalne baze smo uporabili obstoječi slovar SZJ, ki so ga zgradili na Zvezi društev gluhih in naglušnih Slovenije,[1] ob vseh novih kretnjah ali pomenih pa smo leksikalno bazo dopolnili.

Izgovorjava. Artikulacija z glasom ali brez, ki spremlja kretnjo, lahko določa, potrjuje ali spreminja pomen kretnje.

Pomen. Vsaki kretnji je pripisan pomen glede na kontekst besedila.

Sestavljen pomen. Nekatere kretnje so sestavljene iz več delov, denimo DELAVKA iz kretenj DELATI in ŽENSKA. Te so označne v ločenem nivoju.

Grafični zapis v HamNoSys (Schmaling in Hanke 2001). Grafični zapis kretenj pomaga pri določanju različic kretenj in je pomemben korak h generiranju kretenj z animiranimi agenti.

Z označevanjem sta se ukvarjala dva raziskovalca: eden gluh od rojstva in drugi otrok gluhih staršev. Med označevanjem so se pojavljala številna vprašanja o segmentaciji in označevanju sestavljenih kretenj, nejasnih in nedokončanih kretnjah, razlikah med kretnjami in gestikulacijo in še mnoga druga. Reševali smo jih na najboljši možni način in se pogosto posvetovali s sodelavci iz Hamburga, ki se ukvarjajo z označevanjem korpusa nemškega znakovnega jezika DGS.

3. Korpusna analiza leksike SZJ

Razvoj korpusnega jezikoslovja temelji, vse od nastanka elektronskih korpusov v letih 1960 in 1970, na opazovanju in analizi reprezentativnih besedilnih vzorcev. V raziskovanju znakovnih jezikov so kvantitativne metode pomembne še iz enega razloga. Medtem ko pri govorjenih jezikih pojav zapisovanja privede do določene stopnje standardizacije ali vsaj soglasja o obliki zapisanih besed, je pri znakovnih jezikih celotno sporočilo v obliki vizuelne podobe, sestavljene iz gibov rok in telesa, obrazne mimike, artikulacije, gestikulacije in uporabe prostorskih elementov.

Ker znakovni jeziki nimajo standarda za zapisovanje, razen približnih grafičnih zapisov, namenjenih raziskovanju jezika in ne komunikaciji, je proces standardizacije jezika kljub želji mnogih uporabnikov težja naloga. Z analizo korpusnih podatkov si lahko pri standardizaciji SL pomagamo z različicami kretenj in primerjamo njihovo pojavnost.

Poznamo štiri sorodne korpusne raziskave za različne znakovne jezike. Morford in MacFarlane (2003) sta predstavila distribucijsko analizo ameriškega znakovnega

[1] http://www.zveza-gns.si/slovar-slovenskega-znakovnega-jezika/

[2] Ikonična kretnja je zasilni prevod angleškega izraza

jezika ASL z uporabo relativno majhnega korpusa (okoli 4000 kretenj). Veliko večjo zbirko sta uporabila McKee in Kennedy (2006), ki sta raziskala leksikalne značilnosti novozelandskega NZSL z uporabo korpusa Wellington, ki vsebuje več kot 100.000 pojavnic. V zadnjem času sta bili izvedeni še dve raziskavi. Prvo je izvedel Johnston (2011) za avstralski Auslan z uporabo označenega korpusa 63.436 pojavnic, drugo, za britanski BSL, pa Cormier s soavtorji (2011) na korpusu s 24.864 pojavnicami.

Naš pristop k leksikalni analizi je najbolj soroden Johnstonovi raziskavi (prav tam), saj uporabljamo tudi podoben način označevanja. Orodje iLex ima namreč tri pomembne lastnosti, ki omogočajo kvantitativno analizo: prvič, vse kretnje so shranjene v bazi in enolično povezane z glosi. Označevalec vedno izbere glos iz baze, razen če gre za novega, ki ga je potrebno v bazo dodati. Drugič, vsakemu leksemu dodajamo zapis HamNoSys, prav tako se ta zapis doda različicam leksema. Tako je vsak glos mogoče nedvoumno povezati z obliko kretnje. Tretjič pa ima označevalec dostop do video posnetkov vseh pojavitev določenega leksema in s tem možnost medsebojne primerjave za večjo doslednost.

iLex omogoča izvoz podatkov z uporabo ukazov SQL, zato smo za potrebe analize izvozili celotno bazo v Excel. Ker se nekateri deli analize nanašajo na semantične kategorije, ki jih sicer označevalna shema Signor ne vsebuje, smo morali nekatere dele ročno označiti.

3.1. Osnovna statistika korpusa

Celotna velikost označenega korpusa Signor je trenutno (junij 2014) 30.335 pojavnic in 2.976 različnic. 1.043 kretenj se v našem korpusu pojavi le enkrat. Najpogostejši kretnji po frekvenčni listi sta dve različici osebnega zaimka, JAZ1 in JAZ2, ki skupaj predstavljata 3,9 % celotnega zbira podatkov (glej Tabelo 1). Naslednji na spisku je POTEM, ki mu sledi kazalni zaimek TO. Skupna frekvenca prvih 10 kretenj je 10,8 %, kar pomeni, da deset najpogostejših kretenj predstavlja desetino celotnega korpusa. Pri prvih dvajsetih kretnjah je skupna frekvenca 17,4 % in pri stotih 38,9 %.

Če primerjamo naš vzorec SZJ z jezikoma Auslan in BSL, ni opaziti velikih razlik: korpus Auslana ima 55.859 pojavnic in vsebuje 6.171 različnic, od katerih je 3.606 enopojavnic, medtem ko je pri korpusu BSL 24.684 pojavnic z 2.507 različnicami, vendar število enopojavnic ni podano (Cormier 2011). Kazalna kretnja, ki predstavlja zaimek v prvi osebi, je najpogostejša tako v Auslanu kot v BSL-u s skupnima frekvencama 5 % in 6,9 % - v SZJ je frekvenca nekoliko nižja (3,9 %).

Spisek 20 najpogostejših kretenj v korpusu Signor vsebuje le štiri polnopomenske kretnje: DELATI, RAD, LETO in ŠOLA, medtem ko so preostale kretnje kazalne, kot so zaimki (JAZ, JAZ1, TO, MOJ, TAM), in ikonične, ki označujejo smer in/ali obliko.[2] Spisek vsebuje tudi glos za nejasne kretnje, saj v 184 primerih označevalca kljub sobesedilu nista mogla določiti kretnje ali pa sta s to oznako želela opozoriti na primer, o katerem se je potrebno posvetovati.

Višja frekvenca tako imenovanih funkcijskih kretenj, še posebej zaimkov in ikoničnih kretenj, se ujema z ugotovitvami pri Auslan in BSL.

Kot pričakovano je frekvenčni spisek korpusa Signor precej različen od pisane/govorjene slovenščine,[3] kjer se najpogosteje pojavlja pomožni glagol v tretji osebi *je*, ki mu sledijo vezniki (*in, da*), predlogi (*v, na, z, s*), povratnosvojilni zaimek (*se*) in preteklik *biti* (*bil*), prvi polnopomenski element pa se pojavi šele na 21. mestu (*leto*), prav tako je prvoosebni zaimek *jaz* šele na 26. mestu. To razliko lahko razložimo z dejstvom, da je slovenščina za razliko od SZJ tipični sintetični jezik, kjer sta osebni zaimek in povedek združena.

	Glos	**Pogostost**
1	JAZ	687
2	JAZ1	498
3	POTEM	354
4	TO	332
5	DELATI	247
6	PREJ	239
7	A	238
8	IKONIČNO-GIBANJE	229
9	IKONIČNO-OBLIKA	225
10	NE	225
11	JA	221
12	TAKO	218
13	MOJ	208
14	RAD	206
15	TAM	194
16	EN	192
17	NEJASNA KRETNJA	184
18	LETO	182
19	TUDI	177
20	ŠOLA	154

Tabela 1: Prvih 20 najpogostejših kretenj

3.2. Leksikalne in semantične lastnosti

V naslednjih korakih smo želeli raziskati leksikalne in semantične lastnosti besedišča SZJ. Za začetek smo želeli raziskati besednovrstno sestavo besedišča, zato smo na seznamu glosov uporabili oblikoskladenjski označevalnik za slovenščino ToTaLe (Erjavec in dr., 2010). Pri korpusih znakovnih jezikov je tak postopek iz več razlogov problematičen: prvič je glos kretnje zgolj pomenska oznaka, ki predstavlja približno preslikavo pomena v slovenščino, drugič je znano, da v znakovnih jezikih vlogo slovnice prevzemajo popolnoma drugačne

[2] Ikonična kretnja je zasilni prevod angleškega izraza *classifier*, ki se uporablja v tuji literaturi o znakovnih jezikih, pomeni pa poseben semantični razred kretenj, ki imajo atributivno vlogo in nakazujejo obliko, velikost, gibanje ipd. S tem niso mišljene polnopomenske ikonične kretnje, kot je npr. RIBA, ki oponaša gibanje ribe v vodi.

[3] Za primerjavo je uporabljen korpus Gigafida, http://www.gigafida.net.

strukture kot pri govorjenih/pisanih jezikih, in nenazadnje je kretnja s samostalniškim glosom lahko uporabljena v različnih kontekstih kot glagol, samostalnik ali določilo. Poleg tega pri samodejnem besednovrstnem označevanju označujemo osnovne kretnje in ne sestavljenih pomenov, čeprav nekateri sestavljeni samostalniki izhajajo iz glagola, ki mu dodamo kretnjo za osebo ali žensko (UČENEC = UČITI + OSEBA). Iz vseh teh razlogov gre rezultate v Tabeli 2, kjer vidimo distribucijo osnovnih besednih vrst v leksikonu SZJ, razumeti le kot zelo grob približek resničnemu stanju, saj smo pomenske oznake ali glose zgolj preslikali v besedne vrste, kot jih razumemo v slovenščini.

Daleč najpogostejša kategorija je samostalnik, ki ji sledijo glagol, prislov in na koncu pridevnik. Sumimo pa, da bi bilo število samostalnikov še višje, če bi posebej označevali tudi sestavljene kretnje.

Besedna vrsta	Št. različnic
samostalnik	1545
glagol	799
prislov	393
pridevnik	282

Tabela 2: Pogostost besednih vrst

Ročni pregled 300 najpogostejših pojavnic kretenj kaže na pomembnejše teme in semantične skupine, ki jih vsebuje naš korpus. Najpogostejši samostalniki so povezani s časom (LETO, MESEC, KONEC), družino (MAMA, BRAT, SIN), gluhoto (DRUŠTVO, ZAVOD), delom (SLUŽBA) in vsakdanjim življenjem (ŠPORT, FILM, ŠOLA, VRTEC, PRIJATELJ, RAČUNALNIK). Pogosti glagoli so povezani z delom in izobraževanjem (DELATI, UČITI SE, TISKATI, PISATI, ŠTUDIRATI), gibanjem (PRITI, HODITI, PRESELITI, POTOVATI), občutki (SLIŠATI, VIDETI, GLEDATI) in komunikacijo (KRETATI, GOVORITI, POGOVARJATI SE,). Veliko pogostih prislovov je načinovnih (RAD, LEPO, DOBRO, TEŽKO), medtem ko se pridevniki pogosto nanašajo na gluhoto (GLUH, SLIŠEČ, NAGLUŠEN), starost (STAR, MLAD, NOV), lastnost (LEP, VESEL) ali lastnino (MOJ, NJEGOV).

3.2.1. Mašila

Mašila so zanimiva skupina kretenj. Ko smo začeli z označevanjem korpusa Signor, nismo načrtovali ločevanja kretenj na semantične razrede ali pomenske skupine. Vendar pa nas je gluha označevalka kmalu opozorila, da so določene kretnje uporabljane predvsem kot mašila v toku pripovedi, ki pogosto nakazujejo fazo razmišljanja, kako formulirati preostanek izjave.

Skupna frekvenca teh kretenj je 647 pojavnic, kar je 2,1 % našega korpusa. Našli smo 33 različnic kretenj, ki jih lahko obravnavamo kot mašila. V Tabeli 3 je prikazanih deset najpogostejših.

Poglobljene analize mašil in njihove vloge v kretanem besedilu sicer še nismo izvedli, vendar iz naših vzorcev razberemo, da se nekatera mašila uporabljajo kot ločilo med pomenskimi deli kretanega besedila.

	Pogostost
TAKO	218
KAJ PA VEM	82
TO JE VSE	60
KAJ ŠE	55
TAKO JE	35
EH	34
KAJ ČEŠ	30
TA	20
TO JE TO	20
KAKO ŽE	13

Tabela 3: Deset najpogostejših mašil v SZJ

3.2.2. Geste

Cormier in dr. (2011) definirajo geste kot gestam podobne kretnje oziroma nize ponazoritvenih gibov. V korpusu Signor se geste pojavijo v skupni frekvenci 550 pojavnic in predstavljajo 1,8 % celotnega korpusa. Ker geste velikokrat ponazarjajo pomen, za katerega ni primerne kretnje, ali pa se uporabijo za poudarek določenega dogodka, je njihova variabilnost precej velika; preko 130 različnic je opredeljenih kot geste. Nekateri so po pomenu podobni mašilom, vendar je njihova vloga drugačna, spet drugi pa izražajo kompleksnejši pomen, kot npr. [vreči se na tla], [utripajoča luč] ali [sedeti na rami]. Takšni pantomimični gibi se običajno uporabljajo v spontanem kretanju in predstavljajo edinstveno lastnost znakovnega jezika, da je moč kompleksne ali sestavljene pomene izraziti izjemno gospodarno in ekspresivno.

Naša označevalca SZJ ločita med mašili in gestami s poudarkom, da je gesta vedno v funkciji nadomestila kretnje, medtem ko so mašila lahko kombinacija geste in kretnje v funkciji diskurza. Tako je mašilo NE VEM skomig z rameni, ki ga spremljajo navzgor obrnjene dlani, medtem ko gesta KAJ PA VEM predstavlja le skomig z rameni.

3.3. Variacije

Kot pri drugih znanih znakovnih jezikih je tudi v SZJ veliko sinonimov in različic kretenj. Sinonimi so definirani kot raba dveh ali več oblikovno nesorodnih kretenj z enakim pomenom, variacije pa kot raba dveh ali več oblikovno sorodnih kretenj z enakim pomenom. Eden dobro poznanih primerov sinonimije so tri različne kretnje za pomen [zdravnik], vsaka s svojo etimologijo in uporabo v različnih delih Slovenije, medtem ko je primer variacije med kar osmimi zabeleženimi variantami za prvoosebni zaimek JAZ1-JAZ8, kjer je razlika v obliki dlani in mestu telesa, kamor kaže. Žal trenutna označevalna shema korpusa Signor ne beleži razlike med sinonimi in različicami, tako da ne moremo podati kvantitativnih podatkov za posamezni frekvenci teh pojavov.

Od 2.976 različnic je 471 takšnih, ki imajo vsaj en sinonim ali različico, dve kretnji pa jih imata celo osem (JAZ in ITI). Največja variabilnost je pri kretnjah, ki

določajo količino ali obseg nečesa: kretnje za VELIKO, MALO, NIČ, VSE in KONEC. Te kretnje imajo po pet različic.

Pogostost teh različic je pomembna za morebitno leksikografsko obravnavo SZJ, kar je tudi razlog, da smo vsako različico ali sinonim označili z zapisom HamNoSys. Z uporabo tega zapisa lahko prikažemo posamezno kretnjo z animiranim agentom in leksikograf lahko vsako glos poveže z ustrezno kretnjo, ne da bi za to potreboval dostop do korpusa oziroma posnetkov v njem.

4. Zaključek

Projekt Signor predstavlja prvi poskus ustvarjanja označene, reprezentativne in avtentične zbirke besedil v SZJ. Določeni deli označevanja še potekajo, vendar lahko iz podatkov, ki so trenutno na voljo, dobimo prvi vpogled v leksikon SZJ in njegove kvantitativne lastnosti.

Predstavljene številke so do neke mere primerljive s podobnimi raziskavami, ki so bile narejene za BSL, ASL in Auslan, vendar pa je skupna pogostost kazalnih kretenj in gest v SZJ nekaj nižja kot denimo v BSL. Pri tem velja poudariti, da je primerjava pogostosti določenih jezikovnih pojavov med korpusi znakovnih jezikov približno tako nezanesljiva kot primerjava oblikoskladenjskih oznak korpusov dveh jezikov, ki uporabljata različna nabora oznak. Pomembna lekcija, ki smo se je naučili pri označevanju, je, da je razvrščanje kretenj v semantične in slovnične podrazrede subjektivno in zatorej v vsakem primeru pod vplivom osebnega jezikovnega občutka označevalca. Zavedamo se omejitev takega pristopa in v prihodnosti načrtujemo pregled vseh oznak, še posebej ikoničnih kretenj, gest in mašil.

Korpus Signor bo po zaključku primeren za analizo skladenjske strukture SZJ, kar bo predstavljalo tudi temelj za posodobitev gradiva za poučevanje SZJ. V ta namen imamo v načrtu dodajanje novih ravni označevanja, predvsem določanja meje med izjavami. Za poglobljeno analizo leksikalno semantičnih lastnosti načrtujemo poskuse z wordnetovimi sinseti. Razvijamo tudi spletni iskalnik, ki uporabniku omogoča vpogled v rabo posameznih kretenj s sobesedilom. Prek zapisa HamNoSys bo mogoče vsako kretnjo prikazati z animiranim agentom, izseki iz videoposnetkov pa bodo na voljo le za osebe, ki so dovolile objavo posnetkov na spletu.

Dolgoročno bi veljalo razmišljati tudi o razvoju sistema za strojno prevajanje med SZJ in slovenščino; takšni sistemi se že razvijajo za nekatere znakovne jezike (Schmidt in dr., 2011). Korpus Signor bi lahko uporabili kot učno množico za statistična orodja, iz oznak HamNoSys pa je mogoče generirati kretnje z animiranim agentom. Trenutno pa je ovira tudi majhnost korpusa, saj 30.000 pojavnic ni dovolj za izgradnjo jezikovnega in prevodnega modela.

Zahvala

Raziskava je nastala v okviru projekta, ki ga financira ARRS (koda projekta J6-4081). Zahvaljujemo se vsem informantom, ki so sodelovali pri projektu in prispevali posnetke za korpus Signor.

5. Literatura

Cormier, K., J. Fenlon, R. Rentelis in A. Schembri, 2011. Lexical frequency in British Sign Language conversation: A corpus-based approach. *Proceedings of the Conference on Language Documentation and Linguistic Theory 3*, uredili P.K. Austin, O. Bond, L. Marten in D. Nathan. London: School of Oriental and African Studies.

Erjavec, T., D. Fišer, S. Krek, N. Ledinek, 2010. The JOS Linguistically Tagged Corpus of Slovene. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Malta, 2010.

Globačnik, B., 2007. *Stališča Slovencev do slovenskega znakovnega jezika.* Magistrska naloga. Ljubljana: ISH.

Globačnik, B., 2001. Slovenski jezik in slovenski znakovni jezik. Defectologica slovenica 9/2. 19–28.

Hanke, Th. in J. Storz, 2008. iLex – A Database Tool For Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. *Proceedings of the Language Resources and Evaluation Conference 2008*, 28.-30. maj 2008.

Johnston, T., 2011. Lexical frequency in signed languages. *Journal of Deaf Studies and Deaf Education* 17:2. 163-193.

Hanke, Th., S. Matthes, A. Regen in S. Worseck, 2012. Where Does a Sign Start and End? Segmentation of Continuous Signing. In: *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon Language Resources and Evaluation Conference (LREC)* Istanbul, May 2012. 69-74.

Konrad, R., Th. Hanke, S. König, G. Langer, S. Matthes, R. Nishio in A. Regen, 2012. From form to function. A database approach to handle lexicon building and spotting token forms in sign languages. In: *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon Language Resources and Evaluation Conference (LREC)* Istanbul, May 2012. 87-94.

Kuplenik, N., 1999. O jezikovnih napakah pri pisnem izražanju gluhih srednješolcev. *Jezik in slovstvo* 44/5. 43–57.

Logar Berginc, N. in S. Krek, 2010. New Slovene corpora within the "Communication in Slovene" project. *Slavicorp conference.* Warsaw.

Lowenbraun, S., Appelman, K. in Callahan, J., 1980. *Teaching the hearing impaired through total communication*. Columbus, OH: Charles E. Merrill.

McKee, D. in G. Kennedy, 2006. The distribution of signs in New Zealand Sign Language. *Sign Language Studies* 6. 373-390.

Moderndorfer, M., 1989. Totalna komunikacija. V: Z. Juras (ur.): *O tematiki totalne komunikacije in organiziranje gluhih in naglušnih danes in jutri. Zbornik mednarodnega posveta.* Ljubljana: Zveza društev gluhih in naglšnih Slovenije. 123-129.

Morford, J. in J. MacFarlane, 2003. Frequency characteristics of American Sign Language. *Sign Language Studies* 3. 213-225.

Nishio, R., S. Hong, S. König, R. Konrad, G. Langer, Hanke, Th. in Ch. Rathmann, 2010. Elicitation methods in the DGS (German Sign Language) Corpus Project. Poster presented at the *4th Workshop on the Representation and Processing of Sign Languages:*

Corpora and Sign Language Technologies, following the 2010 LREC Conference in Malta, May 22.-23., 2010. 178-185.

Schmaling, C. in Th. Hanke, 2001. HamNoSys 4.0. Dostopno na: http://www.sign-lang.uni-hamburg.de/Projekte/HamNoSys/HNS4.0/englisch/HNS4.pdf.

Schmidt, Ch., D. Stein in H. Ney. 2011. Challenges in Statistical Sign Language Translation. *SLTAT 2011 - International Workshop on Sign Language Translation and Avatar Technology*, Berlin, Germany.

Vintar, Š., B. Jerko in M. Kulovec, 2012. Korpus slovenskega znakovnega jezika. *Zbornik 8. konference Jezikovne tehnologije, ISJT12*. 191-195.

Žele, A., 2007. Kako nevsiljivo povezovati znakovni jezik s pisnim sporočanjem. V: Jasna Bauman (ur.): *Standardizacija slovenskega znakovnega jezika v luči Resolucije o nacionalnem programu za jezikovno politiko 2007-2011*. Zbornik srečanja. Ljubljana: Združenje tolmačev SZJ. 13-17.

Žele, A. in J. Bauman, 2011. Slovenski znakovni jezik med normo in prakso. V: *Meddisciplinarnost v slovenistiki, 30. simpozij Obdobja.* Ljubljana: Univerza v Ljubljani. 557-582.

Variabilnost izgovora kot ovira pri avtomatskem prepoznavanju govora: primer epenteze, epiteze in proteze v govoru slovenskih predšolskih otrok

Martina Ozbič*, Damjana Kogovšek*, Jerneja Novšak Brce*, May Barbara Bernhardt‡, Joseph Stemberger⌘, Mojca Muznik*

* Pedagoška fakulteta Univerza v Ljubljani,
Kardeljeva ploščad 16, 1000 Ljubljana
martina.ozbic@pef.uni-lj.si, jerneja.novsak@pef.uni-lj.si, mojca.zimic@gmail.com
‡ University of British Columbia, School of Audiology & Speech Sciences
Friedman Bldg 2177 Wesbrook Mall, Vancouver BC Canada, V6T 1Z3
⌘University of British Columbia. Totem Field Studios. UBC Department of Linguistics. 2613 West Mall
Vancouver, British Columbia, Canada, V6T 1Z4
may.bernhardt@audiospeech.ubc.ca
joseph.stemberger@ubc.ca

Povzetek

Variabilnost izgovora zaradi različnih narečij, starosti, spola, anatomske strukture govoril, fonološkega statusa, foniranja, resoniranja je ovira za avtomatsko prepoznavanje govora. Še posebno variabilen je izgovor pri otrocih in pri govorcih s komunikacijskimi, jezikovnimi in govornimi težavami. V sklopu mednarodnega projekta o zakasnelem fonološkem razvoju otrok smo analizirali epentetične procese pri 54 predšolskih otrocih in ugotavljali soglasniško in samoglasniško epi-, pro- in epen-tezo . Izkazalo se je, da je pogost proces, ki je značilen v določenih oblikah za odrasle govorce, za mlajše govorce ali pa za osebe s fonološkim zaostankom. Še posebej izstopata epenteza polglasnika v soglasniških sklopih in epiteza zapornika ali pripornika pred priporniki. Iz rezultatov je razvidno, da je samoglasniška epenteza dokaj pogosta v konsonantskih zvezah, še posebej pri tistih, ki zahtevajo različno mesto ali način artikulacije, npr. med obstruenti in sonoranti, med obstruenti in med sonoranti. Konsonantska epenteza se poraja v besedah s težkimi fonemi, še posebej obstruenti pred mehkonebnimi glasovi, pred sičniki in šumniki, pred šumniki. Zanimiv proces dodajanja je tudi dodajanje nezvenečih zapornikov pred pripornike na začetku besede.

Variability of speech as a barrier for automatic speech recognition: epenthesis, prothesis and epithesis in speech of Slovenian preschool children

The variability of speech due to dialects, age, gender, anatomical structures of the speech organs, phonological status, phonation and resonance can be a barrier for automatic speech recognition. Especially in children and among speaker with communication, language and speech difficulties speech is variable. In the project *Cross-linguistic study of protracted phonological (speech) development in children: Slovenian* we have analyzed 54 preschool children and focused our research on epi-, pro- and epen-thesis. The most evident for of process is epenthesis of schwa in consonant clusters and epithesis of a stop or a fricative before fricatives. The data show that vocalic epenthesis frequently occurs in consonant clusters, especially in those clusters which require different manner of articulation, or articulatory organ (or part of it) for example between obstruents and sonorants, between obstruents and between sonorants. Consonantal epenthesis occurs in words with difficult phonemes, especially obstruents, before the velars, before sibilant, before fricatives (adding affricates). An interesting addition is voiceless stop in front of the fricative at the beginning of words.

1. Uvod

Mentalne predstave govora so v bistvu poenostavitve velike govorne variabilnosti govorcev različne starosti, narečne pripadnosti, spola, načina govorjenja, anatomske strukture, miofunkcionalne kontrole, slušnega statusa, resonance, foniranja, artikulacije, koordinacije gibov, verbalnega spomina, morebitnih govorno-jezikovnih težav, motenj. Zaradi navedenega sta že analiza in transkripcija govora zahtevna procesa, ki zahtevata tenkočutno poslušanje in natančni zapis; avtomatsko prepoznavanje govora, ki pa potrebuje dokaj enoznačne ključe, pa je še toliko težje, saj bi morali algoritmi zajeti ne le tipične foneme in alofone ter vse prehode med glasovi govora, temveč še vse prehodne glasove in druge zvočne pojave, ki se zgodijo ob govorjenju, še posebej pri govorcih, ki svoj govor še razvijajo, ali pri govorcih, ki imajo pomembno odstopajoči izgovor, in ki se razlikujejo od govorca do govorca. O tem natančneje Forsberg (2014), kjer poudarja, da je človekovo razumevanje različno od prepoznavanja računalnika, saj uporablja poleg zvočnega signala še vidne signale; človekov uho razlikuje ne-govorne zvoke od govornih, izloči šum; upoštevati moramo variabilnost vseh elementov komunikacijske zanke, nadalje variabilnost zaradi značilnosti govorca, to je starost, spol, hitrost, stil govora, anatomijo vokalskega trakta, socialne in dialektološke vidike. Jufarsky (2000) pravi, da je avtomatsko prepoznavanje govora sistem, ki mora prekodirati akustične signale v zaporedje besed. Skratka, pri zaznavi govora se pri človeku zgodi proces recepcije, percepcije, diskriminacije, kategorizacije/identifikacije, ki sega od čistih fizikalnih pojavov do kognitivnih-jezikovnih, to je prepoznavanja fonemov kot delov zloga, morfemov, besed in nadalje izjav. Za ta namen je torej potrebna natančna analiza in določitev nevariabilnih akustičnih parametrov, ki so sicer v realnosti zelo variabilni. V govoru so glasovi med seboj povezani, kar pomeni, da niso izgovorjeni z

maksimalnim gibom, temveč tekoče prehajajo z enega na drugega. Fonemi tako niso izgovorjeni izolirano, temveč so vpeti v zloge in besede, izjave, ki predstavljajo določeno celoto. Govorimo namreč vedno o večjem številu ravni v skladu z nelinearno fonologijo, in sicer suprasegmentalni in segmentalni, na nivoju razlikovalnih značilnosti, fonema, zloga, naglasa, dolžine besede, melodije, ritma, tempa, more (časovne razdelitve zloga), časa.

Govorci pri govoru morajo upoštevati vse te ravni in ob tem lahko pride do variabilnih realizacij, med temi do vnosa dodatnih glasov - epenteze.

Epenteza, proces, ki je v fokusu članka, pomeni dodajanje zvoka, fonema ali več fonemov v besedo, še posebej v notranje dele besede; gre za pojav, ko zvok dodamo zato, da je izgovor lažji oz. bolj tekoč, še posebej pri težjih zaporedjih ali začetkih oz. koncih besede. Epentezo delimo na ekscrescenco, ko dodamo soglasnik, ter anaptiksis, ko dodamo vokal. Glede na mesto epenteze poznamo paragoge/epitezo in protezo. Epithesis ali paragoge/epiteza je dodajanje fonema na konec besede; po navadi gre za vokal. Prothesis-proteza je dodajanje fonema pred besedo ali zlog in ne vpliva na pomen besede ali strukturo; lahko je vokal ali konsonant.

Epentezo lahko pričakujemo pri mlajših govorcih, še posebej v soglasniških sklopih, pri slednjih tudi v govoru odraslih oseb. V slovenskem jeziku srečujemo namreč zahtevna zaporedja, kot so začetni in končni enodelni, dvodelni, tridelni in štiridelni soglasniški sklopi, kar je izčrpno opisala Srebot Rejec (1990), sledil pa je Unuk (2005). V slovenskem jeziku najdemo besede, ki se začenjajo in končujejo s samoglasniki ali pa s soglasniki; besede lahko vsebujejo različne kombinacije soglasnikov in samoglasnikov, in sicer na začetku ali koncu zloga ter v začetnem, srednjem-ih, končnem zlogu.

Poudariti pa moramo, da se pri mlajšem govorcu poleg procesa epenteze (ki vpliva na strukturo zloga na nivoju kombinacij konzonantov in vokalov, to je C in V) pojavljajo tudi drugi procesi, ki ne načenjajo strukture zloga, temveč spremenijo način ali mesto artikulacije oz. zvenečnost. V kombinaciji z epentetičnimi procesi le-ti močno ogrožajo razumljivost in prepoznavanje sporočila.

2. Namen članka

Cilj našega dela je bil opisati in analizirati epentezo (soglasniško - excrescentia in samoglasniško - anaptyxis) kot fonološki proces pri govorni produkciji slovenskih otrok starih od 3 do 7 let ter ugotoviti uporabo (pojavnost) epenteze v povezanosti s starostjo, zlogovnimi oblikami in v povezavi z mestom in načinom artikulacije fonemov. Ugotoviti smo želeli uporabo (pojavnost) epenteze v povezanosti z zlogovnimi oblikami in v povezavi z mestom in načinom artikulacije fonemov. V sklopu projekta *Cross-linguistic study of protracted phonological (speech) development in children: Slovenian* smo posneli izgovor 54 predšolskih otrok in transkribirali njihov govor z natančno transkripcijo, kjer smo beležili foneme, alofone, diakritike v skladu z IPA (verzija 2005).

3. Metode dela

3.1. Vzorec:

Vzorec je zajemal 3 starostne skupine otrok iz vrtca v centralnem delu Slovenije, in sicer 54 otrok (29 deklic - 54% otrok, 25 dečkov - 46% otrok), v starosti od 3 let in 6 mesecev do 4 let in 5 mesecev: 17 otrok - 32% (11 deklic oz. 20% otrok, 6 dečkov oz. 11% otrok); od 4 let in 6 mesecev do 5 let in 5 mesecev: 25 otrok - 46% (14 deklic oz. 26% otrok, 11 dečkov oz. 20% otrok), od 5 let in 6 mesecev do 6 let in 5 mesecev: 12 otrok - 22% (4 deklice oz. 7% otrok, 8 dečkov oz. 15% otrok). Vzorec je bil delno naključno izbran, saj je bil teritorialno predhodno določen (s čim manjšim vplivom narečij).

3.2. Instrumentarij:

Preizkus, ki smo ga uporabili, je bil oblikovan leta 2010 na Pedagoški fakulteti v sklopu mednarodnega projekta Cross-Linguistic study of proctracted phonological (speech) development in children - prof. J. P. Stemberger, prof. M. B. Bernhardt, dr. M. Ozbič, dr. D. Kogovšek in dr. S. Košir. Preizkus vsebuje 101 besedo različne dolžine: 26 besed je enozložnih, 48 besed je dvozložnih, 20 jih je trizložnih, 7 besed je štirizložnih. Korpus obsega torej 101 x 54 besed, to je 5454 besed. Za natančnejše informacije glejte Muznik (2012) in Marin (2013).

3.3. Postopek zbiranja podatkov:

V vrtcu smo se dogovorili za sodelovanje. Staršem otrok, ki so bili vključeni v raziskavo, smo predstavili cilj in potek raziskave. Starši so s podpisom soglasja privolili v snemanje in pridobivanje podatkov o otrokovem govornem razvoju. Sledila so srečanja z otroki. Snemanje se je odvijalo v najbolj tihi sobi vrtca. Otroku se je razložilo potek snemanja ter njegovo delo. Sledil je uvodni razgovor, da se je otrok sprostil, nam pa je ta pogovor služil za pridobitev globalne slike spontanega, neusmerjenega govora. Sledilo je snemanje in imenovanje posameznih slik (spontano imenovanje je bilo prioritetno, ob težavah smo imenovanje izzvali z zapoznelo imitacijo oz. z neposredno imitacijo). Na koncu smo zopet izzvali spontani govor in omogočili ogled posnetka, če si je otrok to želel. Otrok je pred seboj dobil mapo s slikami, ki jih je moral poimenovati. Material je bil sestavljen iz barvnih fotografij po semantičnih sklopih.

3.4. Programi in tehnična oprema za pripravo posnetkov (ureditev zvočnih in vidnih datotek):

Za pridobivanje baze podatkov govora otrok smo uporabili digitalno kamero Sony Handycam HDR-SR5E z brezžičnim mikrofonom Sony ECM-HW2, ki je bil nameščen približno 15 centimetrov od otrokovih ust, nameščen na za ta namen prirejen brezrokavnik.

Za pretvarjanje iz video v slušne posnetke (v wav format) smo ob obvezni uporabi slušalk uporabljali program VLC. Sledilo je obdelovanje oz. kreiranje zvočnih podatkov tarčnih besed iz celote s programom Cool Edit. Za določanje natančnejših mej med glasovi smo uporabile program Speech Analyzer ali Praat.

3.5. Postopek obdelave podatkov:

Avdio in video posnetke smo najprej pregledali v programu VLC, da smo si oblikovali grob profil izgovarjave otroka. Sledilo je rezanje tarčnih besed s pomočjo programa Cool Edit. Obdelavo posnetkov smo naredili v programu Speech Analyzer (večkratno

poslušanje). Govor smo zapisovali s pomočjo poslušanja ob uporabi slušalk in vidne slike/sonagrama v programu Speech Analyzer ali Praat ter s simboli za IPA 2005, ki smo jih vnesli v Excel-ovo tabelo s pomočjo programov Phon in IPA Assistant ter pisavo Doulos SIL oz. Charis SIL. Sledila je analiza podatkov v Excel-u. Ujemanje transkripcije smo preverili pri 10 otrocih (oz. 19% otrok) med 2 zapisovalkama izgovora. Za analiziranje govora in zapis smo uporabili program Phon – verzija 1.6: program, ki je izdelan za raziskovanja na področju fonološkega razvoja in fonoloških motenj. Ob tem smo upoštevali vse epentetične oblike, ki so osebi, ki je govor transkribirala, predstavljale neobičajen izgovor, čeprav v tipični situaciji takih oblik epenteze lahko ne zaznamo kot moteče (npr. vnos h-ja po velarnih zapornikih).

4. Rezultati

Iz preglednice 1 je razvidno, da so se oblike epentez v različnih fonemskih okolij pojavljanje v različnih frekvencah. Sicer bi morali za korektno analizo računati deleže, vendar je razvidno, da je največ samoglasniških epentez v soglasniških sklopih, največ soglasniških epentez pa pri obstruentih (pripornikih in zapornikih). Soglasniška epenteza je bolj prisotna kot samoglasniška, skupno pa je v 5454 izjavah besed 404 epentetičnih procesov.

SAMOGLASNIŠKA EPENTEZA	**113**
Samoglasniška epenteza pri zapornikih	**10**
Zaporniki: začetni naglašeni zlog	6
Zaporniki: začetni nenaglašeni zlog	2
Zaporniki: končni naglašeni zlog	1
Zaporniki: končni nenaglašeni zlog	1
Samoglasniška epenteza pri zvezah zapornika in zvočnika	**58**
Zapornik + zvočnik: začetni naglašeni zlog	36
Zapornik + zvočnik: začetni nenaglašeni zlog	12
Zapornik + zvočnik: srednji naglašeni zlog	2
Zapornik + zvočnik: srednji nenaglašeni zlog	1
Zapornik + zvočnik: končni nenaglašeni zlog	5
Zapornik+zapornik: začetni nenaglašeni zlog	2
Samoglasniška epenteza pri pripornikih	**8**
Pripornik: začetni naglašeni zlog	6
Pripornik: začetni nenaglašeni zlog	1
Pripornik: končni nenaglašeni zlog	1
Samoglasniška epenteza pri pripornikih v različnih zvezah (+ zapornik, + zapornik+zvočnik, +zvočnik, + pripornik) in zvezi zlitnik + zvočnik	**25**
Pripornik + zapornik: začetni naglašeni zlog	3
Pripornik + zapornik: začetni nenaglašeni zlog	1
Pripornik + zapornik: srednji nenaglašeni zlog	1
Pripornik + zapornik + zvočnik: začetni naglašeni zlog	2
Pripornik+zapornik+zvočnik: začetni nenaglašeni zlog	3
Pripornik+zapornik+zvočnik: končni nenaglašeni zlog	2
Pripornik + zvočnik: začetni naglašeni zlog	10
Pripornik + pripornik: začetni naglašeni zlog	2
Zlitnik + zvočnik: začetni naglašeni zlog	1
Samoglasniška epenteza pri zvočnikih in zvezi zvočnik + zlitnik ter pri samoglasnikih	**16**
Zvočniki: začetni naglašeni zlog	7
Zvočniki: končni nenaglašeni zlog	1
Zvočniki: končni naglašeni zlog	1
Zvočnik + zvočnik: začetni naglašeni zlog	5
Zvočnik + zlitnik: srednji naglašeni zlog	1
Samoglasniki: končni naglašeni zlog	1
SOGLASNIŠKA EPENTEZA	**291**
Soglasniška epenteza pri zapornikih	**54**
Zaporniki: začetni naglašeni zlog	6
Zaporniki: začetni nenaglašeni zlog	4
Zaporniki + zvočniki: začetni naglašeni zlog	34
Zaporniki + zvočniki: začetni nenaglašeni zlog	10
Soglasniška epenteza pri pripornikih ([x], [f], [v], [s], [z]), zvezah in zlitnikih	**50**
Priporniki: začetni naglašeni zlog	12
Priporniki: začetni nenaglašeni zlog	2
Priporniki [s] [z]: začetni naglašeni zlog	32
Zlitniki: začetni naglašeni zlog	3
Zlitniki: začetni nenaglašeni zlog	1
Soglasniška epenteza pri pripornikih ([s] [z]) v zvezi z zvočniki in zaporniki	**50**
Priporniki+zvočniki: začetni naglašeni zlog	23
Priporniki+zvočniki: začetni nenaglašeni zlog	4
Priporniki + zaporniki: začetni naglašeni zlog	17
Priporniki + zaporniki: začetni nenaglašeni zlog	6
Soglasniška epenteza pri pripornikih šumnikih ([ʃ], [ʒ])	**32**
Priporniki [ʃ] in [ʒ]: začetni naglašeni zlog	27
Priporniki [ʃ] in [ʒ]: začetni nenaglašeni zlog	5
Soglasniška epenteza pri zvočnikih v začetnem in srednjem zlogu	**18**
Zvočniki: začetni naglašeni zlog	15
Zvočniki : začetni nenaglašeni zlog	3
Soglasniška epenteza pri samoglasnikih	**40**
Samoglasniki: začetni naglašeni zlog	33
Samoglasniki: začetni nenaglašeni zlog	7
Soglasniška epenteza pri zapornikih v srednjem zlogu	**15**
Zaporniki: srednji naglašeni zlog	3
Zaporniki : srednji nenaglašeni zlog	3
Zvočniki: srednji naglašeni zlog	4
Zvočniki: srednji nenaglašeni zlog	1
Soglasniška epenteza pri zapornikih v končnem zlogu	**3**
Zaporniki: končni naglašeni zlog	3
Zaporniki: končni nenaglašeni zlog	12
Soglasniška epenteza pri pripornikih in zlitnikih v končnem naglašenem zlogu	**11**
Priporniki: končni naglašeni zlog	2
Priporniki: končni nenaglašeni zlog	4
Zlitniki: končni nenaglašeni zlog	5
Soglasniška epenteza pri zvočnikih v končnem zlogu	**18**
Zvočniki: končni naglašeni zlog	9
Zvočniki: končni nenaglašeni zlog	8
Samoglasniki: končni nenaglašeni zlog	1

Preglednica 1: Frekvence epentez (101 besed pri 54 govorcih = 5454 izjav)

4.1. Soglasniška epenteza

Pri zapornikih so v začetnem naglašenem zlogu prisotni vnosi grlnega zapornika, pripornikov ter drsnikov z namenom, da bi omilili prehod z zapornika na samoglasnik. Slišna je tudi aspiracija. Ko pa gre za zveze zapornik+zvočnik, se pojavijo nekateri foni (grlni zapornik, [k]…) kot začetni glasovi za težje foneme, npr. zveze z vibrantom. Tudi prehod z zapornika na nosnik je za otroke težek, zato vnašajo npr. pripornike (npr. [gxn], [kx]). Zlitniki kažejo drugačne smeri procesov epenteze, in sicer dodajanje drsnika med fonema (npr. [tse] v [tsje]), dodajanje zapornika v smislu podvajanja na koncu zloga ([tse] v [tset]) ali pa dodajanje zloga pred samim zlitnikom ([amtʃɛ']). Priporniki pa so se izkazali iz zornega kota soglasniške epenteze kot najbolj zanimivi, saj je očitno dejstvo, da otroci uporabljajo zapornike (ki so lažji, bolj pogosti fonemi, iz fonološkega vidika neoznačeni – *unmarked*, torej lažji, v razvoju zgodnejši). Pojavljajo se epenteze pred pripornikom v obliki zaporniške epenteze ([k], [p]…) , pa tudi priporniške epenteze ([x]), vidno pa je tudi dodajanje zlitnika pred samim pripornikom. Če se

usmerimo na pripornike sičnike, potem je očitno, da se v najbolj pogosti obliki pojavlja epenteza zapornika oz. zlitnika (zvenečega za zveneče pripornike in nezvenečega za nezveneče pripornike), ki se zlijeta in ustvarita zlitnik ali zlitnik z daljšim pripornikom ali pa ostaneta ločena fonema v zaporniško-priporniški zvezi.

Pri pripornikih v zvezi z zvočniki enako kot prej prihaja do epenteze zapornika pred pripornikom v smislu enostavnega dodajanja ali zlitja s pripornikom (rezultat je zlitnik). Sicer se pojavljajo tudi epenteze znotraj zloga po priporniku. Soglasniška epenteza v začetnem naglašenem zlogu sledi podobnim zakonitostim kot doslej za pripornike, in sicer prihaja do dodajanja zapornikov in zlitnikov, pojavlja pa se tudi epenteza po priporniškem sklopu, pa tudi epenteza dodatnih pripornikov, kar porodi neobičajne zveze, npr. [fsx]. Prihaja tudi do odzvenevanja fonemov.

Pri šumnikih se kot epentetični glasovi velikokrat pojavljajo sičniki ali palatalizirane različice. Tudi zaporniki spremenijo šumnik v zvezo zapornik+sičnik (zlitnik). Srečamo tudi vnos – epentezo drsnika, ki mehča realizacijo (npr. pri besedi šola: ['ɕ→ʃo-], [' ɕ ço-], ['fço-].

Pri zvočnikih se epenteze pojavljajo v podobni obliki kot pri samoglasnikih, in sicer s priporniki ([x]) in grlnim zapornikom. Sicer se pojavljajo tudi zaporniške epenteze, npr. [p]. Slednje sledijo mestu artikulacije.

Ko analiziramo epentezo pri samoglasnikih, se v največji meri pojavlja slišen epentetični grlni zapornik ali pa pripornik. Samoglasnik lahko namreč izgovorimo z mehkim prehodom ali pa s trdim in torej z grlnim zapornikom kot začetno namestitev. Dodajanje pripornikov, ki smo jo zabeležili, vsekakor ne sodi v sprejemljivo realizacijo.

Tudi pri zapornikih, kljub temu da so to osnovni glasovi, prihaja do epenteze, in sicer v obliki priporniške epenteze, vnosa zlitnika ustrezne zvenečnosti, vnosa zloga. Zvočniki v srednjem naglašenem zlogu so le za eno deklico bili vzgib za nadomeščanje r-ja s sklopom [dl], kar ni resnična epenteza, temveč rešitev za nadomeščanje oz. aproksimacijo vibranta [r].

Zaporniki v končnem naglašenem zlogu so po navadi tarča priporniške epenteze, medtem ko so epenteze v nenaglašenih zlogih raznolike, npr. epenteza drsnika, dodajanje nosnika, laterala…

Pri pripornikih je v končnem zlogu epenteza po navadi priporniška, prihaja pa tudi do vnosa drsnika ter zapornika. Pri zlitnikih se pojavlja epenteza drsnika (mehčanje), podaljševanje priporniškega dela zlitnika, epenteza zvočnikov.

V končnem naglašenem zlogu se pri zvočnikih realizirajo različne epenteze, in sicer zvočniške, ki rezultirajo v diftonge (npr. [j], [w]) in samoglasniške (zaporniki). Soglasniška epenteza v končnih samoglasnikih je priporniška.

4.2. Samoglasniška epenteza

Zaporniki so prisotni praktično v vseh jezikih sveta; pojavijo se zelo zgodaj, po vokalih in velarnih glasovih dojenčka. Artikulacijsko niso zelo zahtevni, zaznavno pa ponujajo izzive, saj so zelo kratki, tihi in imajo šumno komponento. Opažamo iskanje artikulacijskih rešitev s protezo zloga (V +[m] ali V) na začetku besede ali na koncu besede, epitezo po zaključku artikulacije zapornika. Zaradi težav s koartikulacijo se vokalska epenteza pojavlja tudi znotraj soglasniških sklopov.

Ob stiku zapornika in zvočnika se v velikem številu pojavlja polglasnik, najmanj artikuliran vokal, kot prehod z enega fonema na drug fonem, tako pri zvezah z laterali in vibranti kot tudi z nosniki (npr. [bl], [tr], [gn]), ne glede na zvenečnost zapornika. Pojavlja pa se tudi dodajanje zloga ali glasu [l]. Pri zvezi zapornik+zapornik si zaradi velike stopnje težavnosti otrok pomaga ne le z vnosom polglasnika, temveč še dodatnega zloga. Včasih se otrok zaplete v soglasniških sklopih, kjer jih po eni strani poenostavi ([ščetka]— [ɕte-kəətka]= [ček-tka]), po drugi strani pa še doda soglasnik ([zobna] v [təɕonbna]) ali zlog ([tə] v besedi [zobna] ali [kə] v besedi [ščetka]). Očitne so težave v organizaciji govornih gibov v določenem zaporedju na različnih mestih artikulacije in na različne načine artikulacije.

Pri pripornikih prihaja do vnosa samoglasnikov s posledično diftongizacijo ali do vnosa samoglasnika, po navadi polglasnika z namenom, da bi soglasniški sklop razdelili na dva zloga s strukturo CV in CV. Pri besedah »žoga, žirafa, zobna ščetka« prihaja do zanimivih procesov podaljševanja besede / vnosa zloga pred ali po kritičnem fonemu oz. zaporedju (žirafa v [ʃjəvə'lafa], žoga v ['ʒəlogak] itd.)

Ko se priporniki povezujejo z drugimi fonemi, lahko ugotovimo, da je najpogostejši epentetični glas polglasnik pri zvezi z zapornikom (zveze [zd], [sp], [kl]…), enako velja pri zvezah pripornik+zapornik+zvočnik ([zdr], [str]…), pripornik+zvočnik ([sn]) in pripornik+pripornik ([sv]), srečujemo pa tudi pojave dodajanja zloga (špageti v [ɕpaka'teti]). Pri zvezi zlitnik+zvočnik prav tako prihaja do polglasniške epenteze, enako kot pri pripornikih, saj je zadnji del realizacije zlitnika priporniški. Ne glede na pozicijo oz. naglašenost se epenteza z vnosom polglasnika pojavlja najbolj pogosto.

Pri zvočnikih so zanimivi začetki izgovora, kjer se pri vibrantu in nosniku pojavljajo začetne predpone polglasnika, pri lateralih in nosnikih pa dodajanje zloga na začetku (proteze). Tudi pri zvezi zvočnik+zvočnik prihaja do polglasniške epenteze.

5. Diskusija

Rezultati so pokazali, da je epenteza pogost pojav v izgovoru otrok starih od 3 do 7 let, v obliki ustrezne posledice koartikulacije samoglasniških in soglasniških sklopov ali pa kot kazalnik resnejših težav govora. Samoglasniška epenteza se dosledno pojavlja pri soglasniških sklopih, še posebej pri tistih sklopih, ki zahtevajo drugačno obliko artikulacije, drug artikulacijski organ (ali le njegov del), in sicer med nezvočniki in zvočniki [br], [tr], [dr], [sr], [xr], [kr], [gr], [tl], [vr], [sn] [gn], [gl], [kl], med nezvočniki npr. [sv], [dv] in [sp] in med zvočniki, npr. [ml], [mr]. Iz rezultatov lahko prav tako razberemo, da se samoglasniška epenteza pojavlja v sklopih, kjer se pojavlja različno mesto artikulacije (npr. [tr]), se uporabljajo različni artikulatorji (npr. [br], [vr]…) ali pa se uporablja različna votlina (nosna – ustna, npr. [mr], [ml]). Soglasniška epenteza se pojavlja v besedah s težjimi fonemi, še posebej pri nezvočnikih/obstruentih: pred mehkonebnima [g] in [k] (prihaja do dodajanja grlnega zapornika [ʔ] ali [x] na začetku besede, pred prvim fonemom), pred priporniki (dodajanje zlitnikov), pred mehkonebnim [x] (dodajanje [k]), pred šumnikom

sibilant [ʃ] se doda [s] ali mehčan [ɕ] ali medzobni [ɵ] (drsenje od enega glasu k drugemu); včasih otroci dodajo [n] pred [g]; zanimivo pa je tudi dodajanje [k] ali [f] pred [s] in [p] pred [f] in [x]. Ne glede na pozicijo v besedi ali naglašenost se pri podobnih glasovih pri istih ali podobnih sklopih pojavljajo podobne epenteze; tako prepoznamo epenteze, ki so tipične za zapornike, zvočnike, pripornike ter za posamezne soglasniške ali soglasniško-samoglasniške zveze, npr. aspiracija po zapornikih, glotalni zapornik pred zaporniki in samoglasniki, zlitnik ali dentalni zapornik pred sičniki in šumniki. Zdi se, da otroci nekatere glasove (predvsem tiste, ki so lažji in enostavni za produkcijo) uporabljajo kot začetek za težke foneme, ki se težje artikulirajo in za katere vložijo več napora pri sami izgovarjavi.

Epenteza (še posebej samoglasniška) se velikokrat poraja kot rešitev ob poskusih izgovora soglasniških sklopov (McLeod, van Doorn, Reed, 2001) in sodi med razvojne fonološke procese mlajših otrok, vendar se v našem vzorcu epenteza pojavlja tudi pri starejših otrocih tik pred vstopom v šolo, kar lahko kaže na artikulacijske težave oz. na težave načrtovanja gibov za izvedbo kompleksnih zaporednih hotnih in natančnih gibov govoril za realizacijo govorjenega jezika.

Epenteza v različnih izvedbah epi-, epen- in proteze pa se pojavlja tudi v obliki, ki zaplete samo izgovorno produkcijo, torej poveča število soglasnikov v zaporedju (soglasniška epenteza), kar je razvojno gledano obraten pojav. Po navadi otroci razgradijo kompleksna zaporedja tipa CCV ali CCCV v CV ali zaporedja CV-jev (vokalska epenteza), v primeru soglasniške epenteze pa se raznim zlogovnim strukturam V, CV, VC, CCV, CCCV ipd. doda še dodaten C (soglasnik).

Ob analizi podatkov oz. transkribiranega govora sta se torej pokazali dve vrsti izgovornih rešitev otrok na fonološke probleme: nekateri so rešili izgovorno nalogo s poenostavitvijo zloga (vokalska epenteza), vendar z daljšanjem besede in povečanjem števila zlogov, morebiti tudi s spremembami naglašenega zloga in torej stopice besede (naglasne strukture), drugi otroci pa so skušali ohraniti dolžino besede, število zlogov in naglas, pomagali pa so si z v-stavitvijo/pred-stavitvijo/po-stavitvijo glasu na mesto, kjer jim je bil izgovor težek (nekateri so uporabili vstavitve celotnega zloga, po navadi v zvezi nosnik + samoglasnik). Kljub temu da se zdita obe rešitvi različni, v bistvu nakazujeta različne otrokove rešitve v smislu poenostavitve izgovora: pri vnosu soglasnika se poenostavi (in torej gre na nižjo razvojno stopnjo) zahtevnost strukture zloga in gre za uporabo razvojno zgodnjih vokalov za rešitev fonološkega problema, v primeru konsonantske epenteze pa otrok uporabi razvojno lažji fon ali fonem, da bi deloval kot sprožilec težjega oz. razvojno kasnejšega fonema (npr. epiteza zapornika /t/ pred /s/-jem).

V naboru epentez pa jasno srečamo tudi glasove, ki se vrinejo v govor zaradi koartikulacije (npr. vnos glasu /j/) ali značilnosti govornega aparata (mehčanje sičnikov, šumnikov, vnos glasu /j/). Zanimivo pa je dejstvo, da vsi otroci - tudi z uporabo epentetičnih pojavov – primarno upoštevajo načelo sonornosti oz. strukture zloga po principu zvočnosti (to je: samoglasnik kot jedro, zvočniki ob jedru v začetku ali koncu zloga, fonemi, ki so na skrajni meji zloga pa so obstruenti – zveneči in nezveneči). Tudi epentetični pojavi upoštevajo osnovno in medjezikovno fonotaktično pravilo zvočnosti fonemov v zlogu.

Nekatere oblike epentez srečujemo tudi pri odraslih govorcih, in sicer epiteza grlnega zapornika pred samoglasniki, aspiracija nezvenečih zapornikov, še posebej dlesničnega in mehkonebnega /t/ in /k/, epentezo polglasnika v soglasniške sklope pri počasnem govoru.

Glede na povedano je razvidno, da so nekatere oblike epentez avtomatizmi izreke zaradi pojavov koartikulacije, druge pa so rešitve oseb, ki z artikulacijo imajo težave in se določenih fonemskih zaporedij ne držijo. Iz analize je razvidno, da so oblike epenteze dokaj stabilne in predvidljive (npr. pogost vnos polglasnika med obstruenti in sonoranti ali proteza zapornika pred obstruenti spiranti), kar kljub veliki variabilnosti omogoča predikcijo možnih epentetičnih pojavov in postavitev možnih algoritmov, vsaj na nivoju strukture zloga (zaporedje C in V), težje pa za vsak posamezen fonem/fon. Za avtomatsko prepoznavanje, ki zahteva širok nabor možnih vzorcev govora, so baze podatke različnih izgovorov nujne, še posebej govorcev različne starosti in različne govorne spretnosti. Ob tem pa ne smemo pozabiti, da pri govorcih, kjer se pojavlja veliko število epentetičnih procesov, so po navadi poleg netipičnih struktur zloga prisotni še procesi, ki spreminjajo foneme jezika tako po zvenečnosti, mestu ali načinu artikulacije, kar še v večji meri onemogoča prepoznavanje govorjenega signala.

6. Zaključek

Analiza epenteze je pokazala, da se različni tipi epenteze pri slovensko govorečih otrocih pojavljajo kot odgovor na različno artikulacijsko kompleksno nalogo. Nekatere so artikulacijsko pogojene, druge prehodne oz. razvojne, tretje pa so kazalniki fonoloških težav ali celo motenj.

Epenteza, ki se pri tem pojavi, je lahko torej posledica artikulacijskih zahtev in torej sprejemljiva, lahko pa je izraz otrokovih omejenih fonoloških in izgovornih zmožnosti. Otrokom v predšolskem obdobju soglasniški sklopi predstavljajo velike izzive v izgovarjavi. Do 8. leta starosti pa naj bi oblike epentez, ki jih slušatelj prepozna in sliši, izzvenele z razvojem in zrelostjo pnevmo-fono-artikulacijske kontrole, koordinacije ter fonološkega razvoja. Analiza nam tako nudi vpogled v načine, kako predšolski otroci rešijo fonološki in artikulacijski problem ustreznega izgovora slovenskih besed, še posebej tistih, ki vsebujejo soglasniške sklope in ponuja gradivo za analizo fonološkega razvoja in prikazuje veliko variabilnost izgovora med otroki, med različno starimi otroki. Variabilnost je prisotna tudi pri istem otroku za isto besedo, kar ni redek pojav v razvojnem obdobju.

Strokovnjaki, ki se ukvarjajo z analizo govora za namene opisa ali pa avtomatskega prepoznavanja govora morajo to variabilnost različnih govorcev tako na ravni realizacije posameznih fonemov kot vezave le-teh upoštevati.

7. Literatura

Forsberg, M. *Why is speech recognition difficult*? http://www.speech.kth.se/~rolf/gslt_papers/MarkusForsberg.pdf. pridobljeno: 1.7.2014.

Jurafsky D., Martin J. H. 2000. *Speech and Language Processing. An introduction to Natural Language Processing, Computational Linguistics, and Speech*

Recognition. Prentice Hall, Upper Saddle River, New Jersey.

Marin, A. 2013. *Fonološki (govorni) razvoj otrok med 2;5. in 5;7. letom starosti: transkripcija govora: diplomsko delo.* Ljubljana. Mentorici: Ozbič, Martina. Kogovšek, Damjana. Univ. Ljubljana, Pedagoška fakulteta. http://pefprints.pef.uni-lj.si/id/eprint/1883.

McLeod, S., Van Doorn, J, Reed, V. A. 2001. *Normal acquisition of consonant clusters.* American Journal of Speech - Language Pathology; May 2001; 10, 2. 99-110.

Muznik, M. 2012. *Fonološki razvoj otrok med 3. in 7. letom starosti (transkripcija govora):* diplomsko delo. Ljubljana. Mentorici: Ozbič, Martina. Kogovšek, Damjana. Prešernova nagrada PeF za študijsko leto 2011/2012. Univ. Ljubljana, Pedagoška fakulteta. http://pefprints.pef.uni-lj.si/1287/1/diploma_2.pdf.

Srebot Rejec, T.: Zveze dveh zapornikov v slovenščini in angleščini. *Slavistična revija*, 38 št. 3, 265 – 283. Ljubljana, 1990. Izšlo z meritvami vred v Delovnem poročilu 5819, okt. 1990 pri Inštitutu "Jožef Štefan " v Ljubljani. 265-272, 274 – 276.

Unuk, D. 2005. *Zlog v slovenskem jeziku.* Rokus: Slavistično društvo Slovenije, 2003. serija: Slavistična knjižnica / Rokus, 7.

Končni super pretvorniki za predstavitev slovarjev izgovarjav pri sintezi govora

Žiga Golob*, Jerneja Žganec Gros*, Simon Dobrišek†

*Alpineon d.o.o.
Ljubljana, Slovenija
{ziga.golob, jerneja.gros}@alpineon.si
†Fakulteta za elektrotehniko
Univerza v Ljubljani, Slovenija
simon.dobrisek@fe.uni-lj.si

Povzetek

Končni pretvorniki predstavljajo kompakten način za predstavitev slovarjev izgovarjav, ki jih potrebujemo pri sintezi govora. V članku je predstavljen nov tip končnih pretvornikov, t.i. končni super pretvorniki, s katerimi lahko slovar predstavimo z manjšim številom stanj in prehodov kot s pomočjo minimalnega deterministčnega končnega pretvornika. Končni super pretvornik ohranja deterministčnost, poleg besed iz slovarja pa lahko dodatno sprejme tudi nekatere druge, neznane besede. Pri tem so lahko oddani izhodni alofonski prepisi za določene neznane besede napačni, vendar se izkaže, da je napaka primerljiva s trenutno najboljšimi metodami za določanje grafemsko-alofonske pretvorbe.

Finite-state super transducers for representing pronunciation lexicons in speech synthesis

Finite-state transducers are well suited for compact representations of pronunciation lexicons used in speech synthesis. In this paper, we present a finite-state super transducer, which is a new type of finite state transducer that allows the representation of a pronunciation lexicon with fewer states and transitions than using a conventional minimized and determinized finite-state transducer. A finite-state super transducer is a deterministic transducer that can, in addition to the words comprised in the pronunciation lexicon, accept some other, unknown words as well. The resulting allophone transcription for these words can be false, but we demonstrate that such errors are comparable to the performance of state-of-the-art methods for grapheme-to-phoneme conversion.

1. Uvod

Ključen del pri sintezi govora je sistem za pretvorbo grafemskega zapisa besed v njihov alofonski prepis. Samodejno določanje alofonskega prepisa v slovenščini temelji na množici kontekstno odvisnih pravil, pri čemer moramo poznati besedni naglas (Gros in Mihelič, 1999). Na žalost pa samodejno določanje besednega naglasa slovenskih besed predstavlja težko nalogo (Golob, 2009), zato je za kvalitetno sintezo govora nujna uporaba obsežnih slovarjev izgovarjav.

Slovar izgovarjav predstavlja preslikavo grafemskih zapisov besed v alofonske prepise. Pri pregibno bogatih jezikih, kot je slovenščina, lahko slovarji vsebujejo več milijonov slovarskih vnosov, zaradi česar je lahko njihova uporaba v pomnilniško manj zmogljivih sistemih, kot so npr. vgrajeni sistemi, problematična. V teh primerih je nujna uporaba postopkov, ki omogočajo pomnilniško učinkovito predstavitev slovarjev.

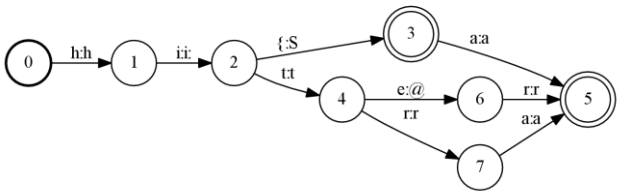
V literaturi je mogoče zaslediti predvsem tri metode, ki omogočajo pomnilniško učinkovito predstavitev slovarjev izgovarjav, in sicer s pomočjo oštevilčenih končnih avtomatov (Lucchesi in Kowaltowski, 1993; Daciuk in Piskorski, 2011), dreves predpon (Ristov, 2005) ter končnih pretvornikov (odslej kratko KP) (Mohri, 1994; Golob at al., 2012). V tem delu bomo predstavili nov način predstavitve s pomočjo končnih super pretvornikov (odslej kratko KSP), ki predstavljajo nekakšno nadgradnjo KP. Poleg manjše predstavitve slovarjev v primerjavi s KP, lahko s KSP z visoko točnostjo določimo alofonski prepis tudi nekaterim neznanim besedam oz. besedam, ki niso vsebovane v izvirnem slovarju.

V članku bomo najprej na kratko predstavili KP ter prikazali, kako lahko z njimi predstavimo slovar izgovarjav. Nadalje bomo pokazali, da zastopanost pregibnih oblik v slovarju močno vpliva na velikost KP. Sledila bo predstavitev t.i. KSP, ki predstavljajo nov način predstavitve slovarjev, nazadnje pa bomo podali še rezultate predstavitve slovarja s KSP ter ocenili napako, ki jo naredimo, če s KSP poskušamo narediti grafemsko-alofonsko pretvorbo za besede, ki niso del slovarja.

1.1. KP ter predstavitev slovarja izgovarjav

KP sestavljajo stanja ter prehodi med stanji. Vsak prehod ima vhodno in izhodno oznako. Ko se na vhodu KP pojavi določen vhodni niz, se ta nahaja v začetnem stanju. KP nato po vrsti sprejema vhodne simbole. Pri vsakem sprejetju vhodnega simbola odda izhodni niz simbolov, ki ga določa izhodna oznaka pripadajočega prehoda, ter se premakne v naslednje stanje. Če za poljuben vhodni simbol v trenutnem stanju ne obstaja prehod, ki ima vhodno oznako enako temu simbolu, pravimo, da KP vhodnega niza ne sprejema. Če se KP po prejetju vseh simbolov vhodnega niza nahaja v končnem stanju, pravimo, da vhodni niz sprejema, pri tem pa postane oddan izhodni niz veljaven. Omenimo še to, da je lahko vhodna ali/in izhodna oznaka enaka praznemu simbolu oziroma nizu.

KP, ki imajo v poljubnem stanju največ en prehod z določeno vhodno oznako, pravimo deterministični KP. Za takšne KP je hitrost pretvorbe vhodnega niza v izhodni niz zelo hitra in ob primerni izvedbi odvisna samo od dolžine vhodnega niza. Druga prednost determinističnih KP je ta, da obstajajo učinkoviti algoritmi za njihovo minimizacijo. Tako dobimo minimalni KP, ki ima najmanjše število prehodov in stanj med vsemi ekvivalentnimi KP (Mohri, 1997), torej KP, ki za poljuben sprejet vhodni niz oddajo enak izhodni niz.

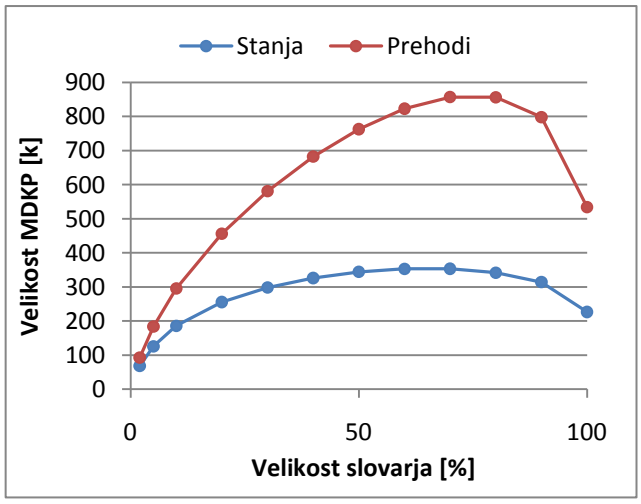


Slika 1: Primer KP, ki predstavlja slovar izgovarjav za tri slovenske besede. Krogi predstavljajo stanja, puščice pa prehode med stanji. Vsak prehod je označen z vhodno in izhodno oznako, ki sta ločeni z dvopičjem. Začetno stanje je označeno z odebeljenim krogom, končna stanja pa z dvojnim krogom.

Vseh KP ni mogoče determinizirati, saj imajo deterministični KP manjšo izrazno moč kot nedeterministični (Hellis, 2004). KP, ki predstavlja slovar izgovarjav, lahko vedno determiniziramo, če iz slovarja odstranimo enakopisnice. Slika 1 prikazuje primer minimiziranega in determiniziranega KP (odslej kratko MDKP), ki predstavlja slovar za štiri slovenske besede.

2. Vpliv velikosti slovarja izgovarjav na velikost KP

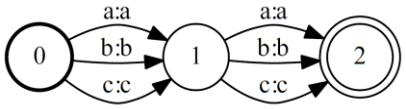
V tem eksperimentu smo želeli preveriti odvisnost velikosti KP od velikosti slovarja, ki ga želimo predstaviti. Na voljo smo imeli slovar SI-PRON za slovenski jezik, ki vsebuje več kot milijon različnih slovarskih vnosov (Žganec-Gros, Cvetko-Orešnik, Jakopin, 2006). Z naključnim izbiranjem slovarskih vnosov smo zgradili 11 pod-slovarjev različnih velikosti in za vse pod-slovarje zgradili MDKP. Rezultate števila stanj in prehodov pridobljenih MDKP prikazuje graf 1.



Graf 1: Odvisnost velikosti MDKP od velikosti slovarja izgovarjav, pri čemer so vnosi v slovar izbrani naključno iz prvotnega slovarja. Opazimo lahko obrat trenda rasti števila stanj in prehodov MDKP.

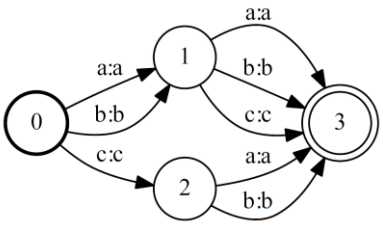
Iz rezultatov lahko razberemo, da velikost MDKP doseže vrh pri 70% do 80% velikosti prvotnega slovarja. Z drugimi besedami, velikost MDKP začne pri določeni velikosti z dodajanjem novih besed oz. slovarskih vnosov iz slovarja padati.

Da bi si ta pojav lahko lažje predstavljali, poglejmo minimalni primer, ki prikazuje mehanizem tega zmanjšanja velikosti MDKP. Kot primer vzemimo izmišljen slovar, katerega ključi[1] so sestavljeni iz vseh možnih izborov dveh črk od treh možnih, npr. črk *a*,*b* in *c*. Na ta način dobimo 9 različnih ključev, in sicer: *aa*, *ab*, *ac*, *ba, bb*… Zaradi enostavnosti naj bodo pripadajoče vrednosti enake ključem. MDKP za ta slovar prikazuje slika 2.



Slika 2: MDKP za izmišljen slovar, katerega ključi so sestavljeni iz vseh možnih izborov dveh črk od treh možnih – *a*, *b* in *c*. Pri tem so vrednosti enake ključem.

Sedaj iz našega izmišljenega slovarja odstranimo slovarski vnos *cc : cc* ter ponovno zgradimo MDKP. Rezultat prikazuje slika 3.



Slika 3: MDKP za enak slovar, kot ga predstavlja MDKP na sliki 2, pri čemer mu manjka slovarski vnos *cc : cc*.

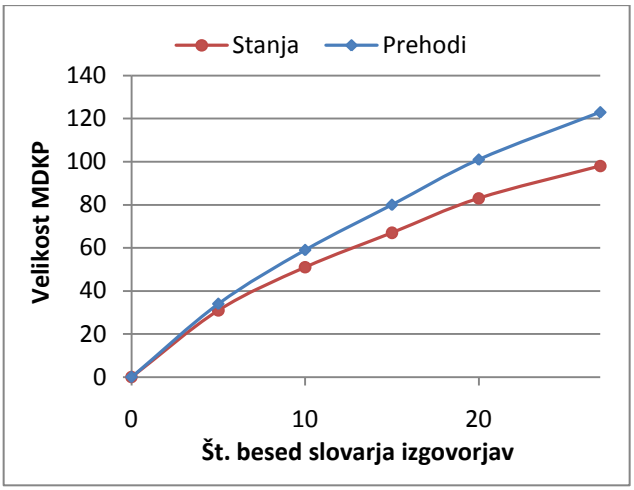
Opazimo lahko, da se je pri odstranitvi slovarskega vnosa iz slovarja kompleksnost MDKP povečala, saj je za predstavitev slovarja potrebno eno dodatno stanje ter dva dodatna prehoda.

V naslednjih poglavjih bomo podrobneje raziskali vzroke, ki vplivajo na zmanjšanje MDKP pri predstavitvi slovarja pri dodajanju novih slovarskih vnosov v slovar.

2.1. Vpliv množičnosti pregibnih oblik na velikost slovarja izgovarjav

Preverili smo vpliv množičnosti pregibnih oblik lem besed iz slovarja na velikost MDKP. Pri tem z množičnostjo pregibnih oblik mislimo na število različnih pregibnih oblik za določeno lemo. Za primer smo vzeli besedo *skopati* ter v slovarju poiskali vse slovarske vnose, katerih grafemski zapisi predstavljajo pregibne oblike leme izbrane besede. Dobili smo 27 različnih slovarskih vnosov, iz katerih smo s pomočjo naključnega izbiranja vnosov tvorili še štiri različno velike pod-slovarje. Za vsak pod-slovar smo zgradili MDKP. Rezultate prikazuje graf 2.

[1] Slovarski vnosi so sestavljeni iz para ključ, vrednost. Pri slovarju izgovarjav tako grafemski zapis predstavlja ključ, alofonski prepis pa vrednost.



Graf 2: Odvisnost velikosti MDKP od števila besed v slovarju izgovarjav. Vse besede slovarja pripadajo isti lemi.

Iz rezultatov je razvidno, da hitrost naraščanja velikosti MDKP z večanjem slovarja rahlo pada, vendar pa ni opaziti obrata trenda povečevanja MDKP.

2.2. Vpliv zastopanosti pregibnih oblik na velikost slovarja izgovarjav

Poglejmo sedaj, kako na velikost MDKP vpliva zastopanost pregibnih oblik v slovarju, sestavljenem iz večih besed, ki se podobno pregibajo. Iz slovarja SI-PRON smo izbrali 28 grafemskih zapisov besed, katerih pregibne oblike imajo 9 različnih končnic ter pripadajo štirim različnim lemam - *potop, osmod, zasp, natoč*. Izbrane leme ter pripadajoče končnice so prikazane v tabeli 1. Lema *zasp* pri tem predstavlja izjemo, ki se pregiba nekoliko drugače kot ostale tri.

MOŽNE LEME	MOŽNE KONČNICE
potop, osmod, zasp, natoč	iš,im,imo,ite,ijo
potop, osmod,natoč	i+l,i+li
zasp	a+l, a+li

Tabela 1: Tabela prikazuje postopek za tvorjenje vseh besed, ki so vsebovane v slovarju. V levem stolpcu so navedene leme besed, v desnem pa možne končnice.

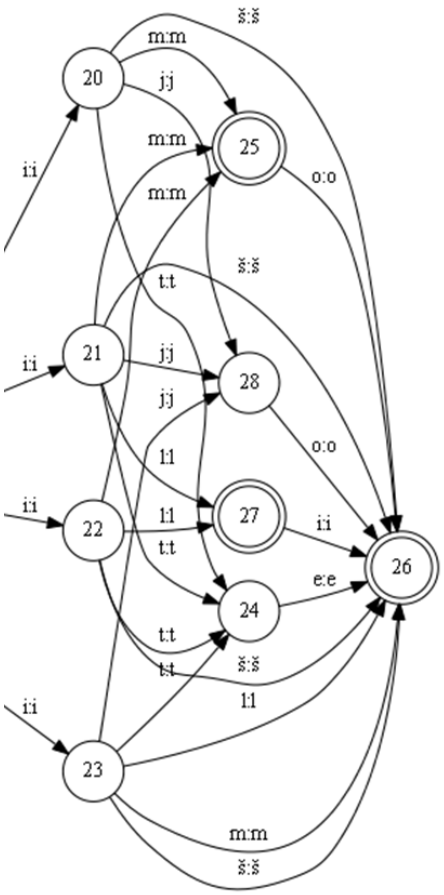
Iz teh besed smo nato tvorili slovar, pri čemer smo zaradi enostavnosti vrednosti ključev izenačili s ključi. Nato smo z naključnim izbiranjem iz tega slovarja tvorili še štiri različno velike pod-slovarje. Za vse tako zgrajene slovarje smo nato zgradili MDKP. Rezultate prikazuje tabela 2.

ŠT. BESED	ŠTEVILO STANJ	ŠTEVILO PREHODOV
5	26	29
9	29	35
17	29	40
23	29	45
28	26	36

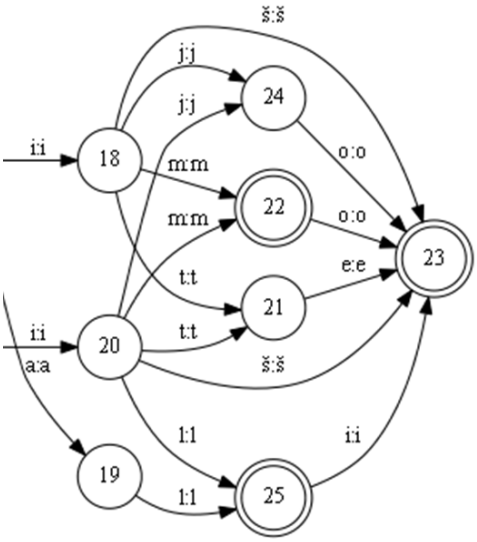
Tabela 2: Tabela prikazuje rezultate števila stanj in prehodov MDKP za vse velikosti pod-slovarjev.

Iz tabele je razvidno, da je velikost MDKP, ki predstavlja vseh 28 vnosov slovarja, manjša od MDKP, ki predstavljata slovarja s 23 in 17 vnosi, število stanj pa je večje celo pri MDKP, ki predstavlja slovar z 9 vnosi. Rezultati nakazujejo, da zastopanost pregibnih oblik močno vpliva na kompleksnost pridobljenega MDKP ter lahko vpliva na obrat trenda rasti velikosti MDKP.

Sliki 4 in 5 prikazujeta shematski prikaz dela MDKP, ki predstavlja končnice besed, pri čemer slika 4 pripada MDKP za slovar s 23 vnosi, slika 5 pa MDKP za slovar z 28 vnosi. Razvidno je, da je kompleksnost MDKP, ki predstavlja slovar s 23 vnosi, precej večja.



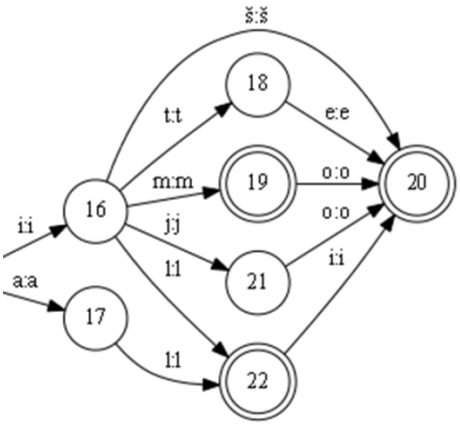
Slika 4: Del MDKP, ki predstavlja slovar s 23 vnosi. Prikazan je le del, ki pretvarja končnice vnosov.



Slika 5: Del MDKP, ki predstavlja celoten slovar z vsemi 28 vnosi. Prikazan je le del, ki pretvarja končnice vnosov.

Smiselno je torej, da so v slovarju, ki ga želimo realizirati s KP, prisotne vse možne pregibne oblike, saj si lahko v tem primeru leme, ki se enako pregibajo, del končnega pretvornika, ki pretvarja končnice, v celoti delijo. Kompleksnost pri tem še vedno povečujejo besede oz. leme besed, ki imajo med pregibnimi oblikami kakšno izjemo, ki se pregiba nekoliko drugače. V našem izmišljenem slovarju je to lema *zasp*, katere dve pregibni obliki imata nekoliko drugačno končnico, in sicer končnico *al* ter *ali* namesto *il* ter *ili*.

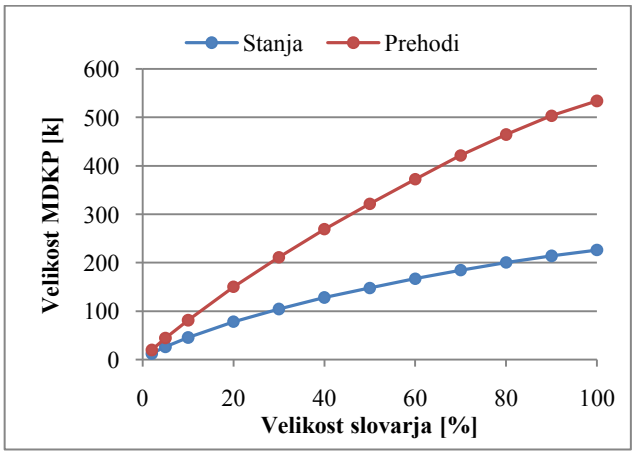
MDKP sprejme samo vnose, ki so vsebovani v slovarju. Če za določeno aplikacijo tako stroga zahteva ni potrebna in je dovolj, da MDKP sprejme vse vnose iz slovarja, ga lahko naprej poenostavimo. Še enostavnejšo obliko bi namreč dobili, če bi za vse štiri leme iz slovarja obstajale pregibne oblike za vseh 9 možnih končnic. V slovar lahko tako dodamo dodatne vnose in sicer vnose z lemami *potop*, *osmod*, *natoč* ter končnicama *al* ter *ali*, ter vnosa z lemo *zasp* in končnicama *il* ter *ili*. Pridobljeni slovar ima tako 36 vnosov, MDKP pa se poenostavi na 23 stanj in 30 prehodov. Shema dela MDKP, ki pretvarja končnice, je prikazana na sliki 6.



Slika 6: Del MDKP, ki predstavlja slovar izgovarjav s 36 vnosi. Prikazan je le del, ki pretvarja končnice vnosov.

Struktura MDKP, prikazana na sliki 6, ki predstavlja slovar z 28 prvotnimi vnosi ter dodatnimi 8 vnosi je torej še nekoliko bolj enostavna kot struktura MDKP, ki predstavlja slovar z le 28 prvotnimi vnosi. Z dodatnimi vnosi smo torej poenostavili strukturo MDKP. S pomočjo KSP, ki ga bomo predstavili v naslednjem poglavju, bomo to idejo posplošili.

Eksperimenti, ki nakazujejo, da na obrat trenda rasti MDKP vpliva predvsem zastopanost pregibnih oblik, so bili izvedeni na poenostavljenem slovarju, katerega vrednosti so bile enake ključem. Da bi pokazali, da podobno velja tudi v primeru dejanskih slovarjev izgovarjav, smo iz slovarja izgovarjav SI-PRON ponovno tvorili 11 različno velikih pod-slovarjev z naključnim izbiranjem, vendar pa smo v tem primeru naključno izbirali le leme besed, nato pa smo vključili še vse pripadajoče pregibne oblike. Za vse pod-slovarje smo nato zgradili MDKP. Rezultate prikazuje graf 3.



Graf 3: Odvisnost velikosti MDKP od velikosti slovarja izgovarjav, pri čemer so v slovarju vedno vsebovane vse pregibne oblike. Povečevanje MDKP je tokrat skoraj linearno odvisno od števila vnosov v slovarju. Opaziti je le rahlo upadanje trenda rasti.

Vidimo lahko, da tokrat ne pride do obrata trenda rasti, kar potrjuje našo hipotezo.

3. Končni super pretvornik (KSP)

V prejšnjem poglavju smo pokazali, da lahko s pomočjo dodatnih, izbranih slovarskih vnosov v slovar zmanjšamo kompleksnost MDKP. Problem predstavlja iskanje takšnih slovarskih vnosov, ki bi zmanjšali kompleksnost, še posebej v primeru realnih slovarjev, kot so npr. slovarji izgovarjav, ki so prvič večji, drugič pa se ključ in vrednost posameznih slovarskih vnosov razlikujeta, s čimer je iskanje primernih slovarskih vnosov težja naloga. Problema smo se zato lotili na drugačen način, in sicer tako, da smo združevali določena stanja, pri čemer smo želeli zadostiti naslednjima dvema pogojema:

- Pridobljen KP mora ostati deterministične.
- Pridobljen KP mora sprejemati vse ključe prvotnega slovarja ter za sprejete ključe oddati pravilne pripadajoče vrednosti.

Tako smo lahko združevali samo stanja, ki so imela določene lastnosti. Takšna stanja smo poimenovali

združljiva stanja. Dve stanji sta združljivi, če zadoščata naslednjim pogojem.

- Če je eno od stanj končno stanje, stanji ne smeta imeti izhodnih prehodov s praznimi vhodnimi simboli oz. ε simboli. Rezultat združevanja takšnih stanj je lahko nedeterministični KP.
- Stanji nimata izhodnih prehodov z enakimi vhodnimi simboli ter različnimi izhodnimi simboli.
- Stanji nimata izhodnih prehodov z enakimi vhodnimi simboli ter enakimi izhodnimi simboli, ki prehajajo v različna naslednja stanja, ki so nezdružljiva.

Da bi lahko določili združljiva stanja, je potrebno preveriti zgornje pogoje, kar pa je v praksi lahko problematično, saj je preverjanje združljivosti stanj zaradi rekurzivnosti, ki je lahko ciklična, zahtevno. V ta namen smo zadnji pogoj poenostavili:

- Stanji nimata izhodnih prehodov z enakimi vhodnimi simboli ter enakimi izhodnimi simboli, ki prehajajo v različna naslednja stanja.

Zaradi poenostavitve pogoja za združljivost stanj nekaterih združljivih stanj nismo mogli zaznati.

KSP smo zgradili tako, da smo najprej zgradili MDKP, nato pa smo nadalje združili vsa stanja, ki so združljiva. Za vsako stanje je bilo potrebno preveriti, ali je združljivo s katerim koli drugim stanjem. Ker nekatera stanja postanejo združljiva šele, ko združimo neka druga stanja, je bilo potrebno to storiti v več iteracijah.

4. Predstavitev slovarja izgovarjav s KSP

Za slovar izgovarjav SI-PRON smo najprej zgradili MDKP s pomočjo odprtokodnega orodja OpenFST (Cyril at al., 2007), nato pa smo s postopkom, ki smo ga opisali v poglavju 3, zgradili še KSP. Tabela 3 prikazuje število stanj in prehodov MDKP in KSP.

		MDKP	KSP	Zmanj.
En izhodni simbol	Stanja	226.363	172.833	23.6%
	Prehodi	534.061	428.114	19.8%

Tabela 3: Zmanjšanje števila stanj in prehodov pri gradnji KSP iz MDKP.

Opazimo lahko, da smo velikost MDKP uspeli zmanjšati za približno 20%.

Čeprav lahko s KSP vnose v slovarju predstavimo z manjšim KP kot v primeru MDKP, pri tem izgubimo informacijo o tem, katere besede so vsebovane v slovarju. Tako se lahko zgodi, da KSP sprejme določeno besedo, ki je slovnično pravilna, vendar ni bila vsebovana v slovarju. V tem primeru je lahko oddan alofonski prepis napačen. V naslednjem poglavju smo poskušali oceniti napako, ki jo naredimo, če za predstavitev vnosov slovarja namesto MDKP uporabimo KSP.

4.1. Ocena verjetnosti napake KSP pri predstavitvi slovarja izgovarjav SI-PRON

Ker pri uporabi KSP izgubimo informacijo o tem, ali je določena beseda vsebovana v slovarju, lahko besede, ki niso del slovarja, pretvorimo napačno v njihov alofonski prepis. Če bi takšno informacijo imeli, bi lahko takšne besede namesto s KSP v alofonski prepis pretvorili s pomočjo kakšnih drugih metod, npr. s pomočjo metod strojnega učenja ali kakšnih drugih statističnih metod.

Da bi ocenili verjetnost napake, ki jo na ta način naredimo, smo slovar SI-PRON naključno razdelili v dva pod-slovarja, pri čemer je prvi vseboval 90% besed, drugi pa preostalih 10% besed in je služil kot testni del. Za prvi del smo zgradili MDKP ter KSP. Rezultate gradnje prikazuje tabela 4.

	MDKP	KSP	Zmanjšanje
Stanja	315.191	190.842	39.5%
Prehodi	800.026	478.315	40.2%

Tabela 4: Rezultati gradnje MDKP in KSP za pod-slovar, ki je vseboval 90% vnosov slovarja izgovarjav SI-PRON.

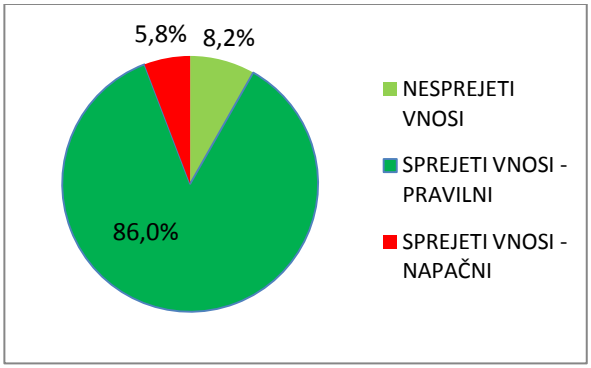
Opazimo lahko, da je tokrat zmanjšanje števila stanj in prehodov precej večje kot v primeru celotnega slovarja in je glede na MDKP približno 40%. Poleg tega je končno število stanj in prehodov manjše kot v primeru gradnje MDKP za celoten slovar. Iz tega lahko sklepamo, da je gradnja KSP še posebej smiselna, ko v slovarju niso vsebovane vse pregibne oblike.

Nadalje smo grafemske zapise 124.099 slovarskih vnosov iz testnega dela slovarja dali na vhod zgrajenega KSP. Za grafemske zapise, ki jih je KSP sprejel, smo spremljali, če se pri tem oddan alofonski prepis ujema z alofonskim prepisom pripadajočega slovarskega vnosa. Rezultate prikazuje tabela 5.

Št. vseh testnih vnosov	124.099
Nesprejeti testni vnosi	10.190
Sprejeti testni vnosi (pravilni)	106.698
Sprejeti testni vnosi (napačni)	7.211

Tabela 5: Tabela prikazuje število sprejetih ter nesprejetih grafemskih zapisov slovarskih vnosov testnega slovarja, ko te damo na vhod KSP. Pravilne sprejete vnose predstavljajo tisti slovarski vnosi, pri katerih je KSP oddal pravilen alofonski prepis.

Bolj pregledno razmerja med posameznimi skupinami slovarskih vnosov prikazuje graf 4.



Graf 4: Razmerja med nesprejetimi ter sprejetimi vnosi, med katerimi nadalje ločimo tiste, za katere oddan alofonski prepis je bil bodisi pravilen bodisi napačen.

Iz rezultatov lahko razberemo, da je odstotek sprejetih vnosov kar 91.8%. Pri tem naj še enkrat opozorimo, da ti vnosi niso bili vsebovani v slovarju izgovarjav, iz katerega smo zgradili KSP. Od sprejetih vnosov je delež tistih, za katere je KSP oddal napačen alofonski prepis, le 6.7%.

5. Zaključek

V članku je predstavljen nov tip KP, ki smo jih poimenovali končni super pretvorniki (KSP), ki poleg želenih besed sprejemajo še nekatere druge z namenom, da lahko pretvorbo želenih besed predstavimo bolj kompaktno.

Pokazali smo, da lahko pri predstavitvi slovarja izgovarjav s pomočjo KSP število stanj in prehodov zmanjšamo za približno 20%, ko so za vsebovane leme v slovarju izgovarjav prisotne tudi vse pripadajoče pregibne oblike besed oz. kar 40% v primeru, ko vse pregibne oblike niso vsebovane.

Ker KSP sprejemajo še druge, neznane besede, za katere lahko oddajo napačen izhodni niz, so KSP uporabni predvsem v aplikacijah, kje ne potrebujemo informacije o tem, katere besede so vsebovane v KP ampak le informacijo o pravilni pretvorbi danih besed oz. besed, iz katerih smo zgradili KSP.

Za slovarje izgovarjav, ki jih uporabljamo pri sintezi govora, si želimo, da pokrivajo čim večji delež besed, saj omogočajo najvišjo stopnjo točnosti pri pretvorbi v alofonski prepis. Vseeno, razen v primeru zaprtih domen, ne morejo vsebovati vseh besed, ki se lahko pojavijo, saj se jezik nenehno spreminja in pri tem stalno nastajajo nove besede. Tako lahko pri uporabi KSP za predstavitev slovarja izgovarjav pride do napake pri pretvorbi v alofonski prepis, ko se na vhodu pojavi neznana beseda. Pokazali smo, da je ta napaka razmeroma majhna, saj je bilo za naš testni slovar od več kot 90% sprejetih besed napačno pretvorjenih le 5.8%. Vidimo torej, da lahko KSP uporabimo tudi kot prepoznavalnik za določanje alofonskega prepisa neznanim besedam, pri čemer je njegova napaka le 6.7%. Tako nizka napaka pa je primerljiva oz. celo manjša od napake namenskih prepoznavalnikov, kjer je ta za slovenski jezik odvisna predvsem od točnosti napovedovanja naglasnega mesta in se giblje nekoliko nad 15% (Golob, 2009).

Pri ocenjevanju verjetnosti napake KSP je bil prvotni slovar SI-PRON naključno razdeljen na testno in učno množico. Tako so bile pregibne oblike besed za določene leme lahko vsebovane tako v učni kot v testni množici. Ker so si pregibne oblike, ki pripadajo isti lemi, med seboj precej podobne, je napovedovanje alofonskega prepisa takšnim besedam iz testne množice, ki so vsebovane tudi v učni množici, nekoliko lažja naloga. V nadaljnjih poskusih bi bilo zato potrebno prvotni slovar razdeliti na učno in testno množico tako, da bi se naključno izbiralo le leme besed, nato pa bi se poleg vključile še vse pripadajoče pregibne oblike. V tem primeru pričakujemo, da bi bila napaka pri pretvorbi neznanih besed nekoliko višja.

6. Zahvala

Raziskovalno delo prvega avtorja je delno financirala Evropska unija iz evropskega socialnega sklada ter sklada za regionalni razvoj v okviru Operativnega programa krepitve regionalnih razvojnih potencialov za obdobje 2007 do 2013, po pogodbi št. P-MR-10/94.

7. Reference

Cyril A., Michael R., Johan S., Wojciech S., Mohri M., 2007. OpenFst: A General and Efficient Weighted Finite-State Transducer Library. Proceedings of the 12th International Conference on Implementation and Application of Automata (CIAA 2007). Lecture Notes in Computer Science, Prague, Springer-Verlag, Heidelberg, Germany, 4783: 11-23.

Daciuk J., Piskorski J., Ristov S., 2011. Natural Language Dictionaries Implemented as Finite Automata. Scientific Applications of Language Methods. London : Imperial College Press, World Scientific Publishing.

Golob Ž., 2009. Samodejno določanje mesta besednega naglasa pri sintezi slovenskega govora. Diplomsko delo, fakulteta za elektrotehniko v Ljubljani.

Golob Ž., Žganec-Gros J., Žganec M., Vesnicer B., Dobrišek S., 2012. FST-Based Pronunciation Lexicon Compression for Speech Engines. *International Journal of advanced robotic systems,* 9: 2011.

Gros J., Mihelič F., 1999. Acquisition of an Extensive Rule Set for Slovene Grapheme-to-Allophone Transcription. Proceedings 6th European Conference on Speech Communication and Technology. September 5–9. 1999. Eurospeech 1999. Budapest, 5: 2075–2078.

Hellis T., 2004. On minimality and size reduction of one-tape and multitape finite automata. Doktorska disertacija.

Lucchesi C., Kowaltowski T., 1993. Applications of Finite Automata Representing Large Vocabularies. *Software-Practice & Experience,* 23: 15-30.

Mohri M., 1994. Compact Representations by Finite-State Transducers. 32nd Meeting of the Association for Computational Linguistics (ACL '94). Proceedings of the Conference. Las Cruces. NM, pp. 204–209.

Mohri M., 1997. Finite-State Transducers in Language and Speech Processing. *Computational Linguistics*, 33: 269–311.

Ristov S., 2005. LZ Trie and Dictionary Compression. Jurnual Software-Practice & Experience, pp. 445–465.

Žganec-Gros J., Cvetko-Orešnik V., Jakopin P., 2006. SI-Pron Pronunciation Lexicon: A New Language Resource for Slovenian. Informatica, 30: 447–452.

Indeks avtorjev / Author index

Konferenca / Conference
Uredili / Edited by

Jezikovne tehnologije /
Language Technologies
Tomaž Erjavec, Jerneja Žganec Gros