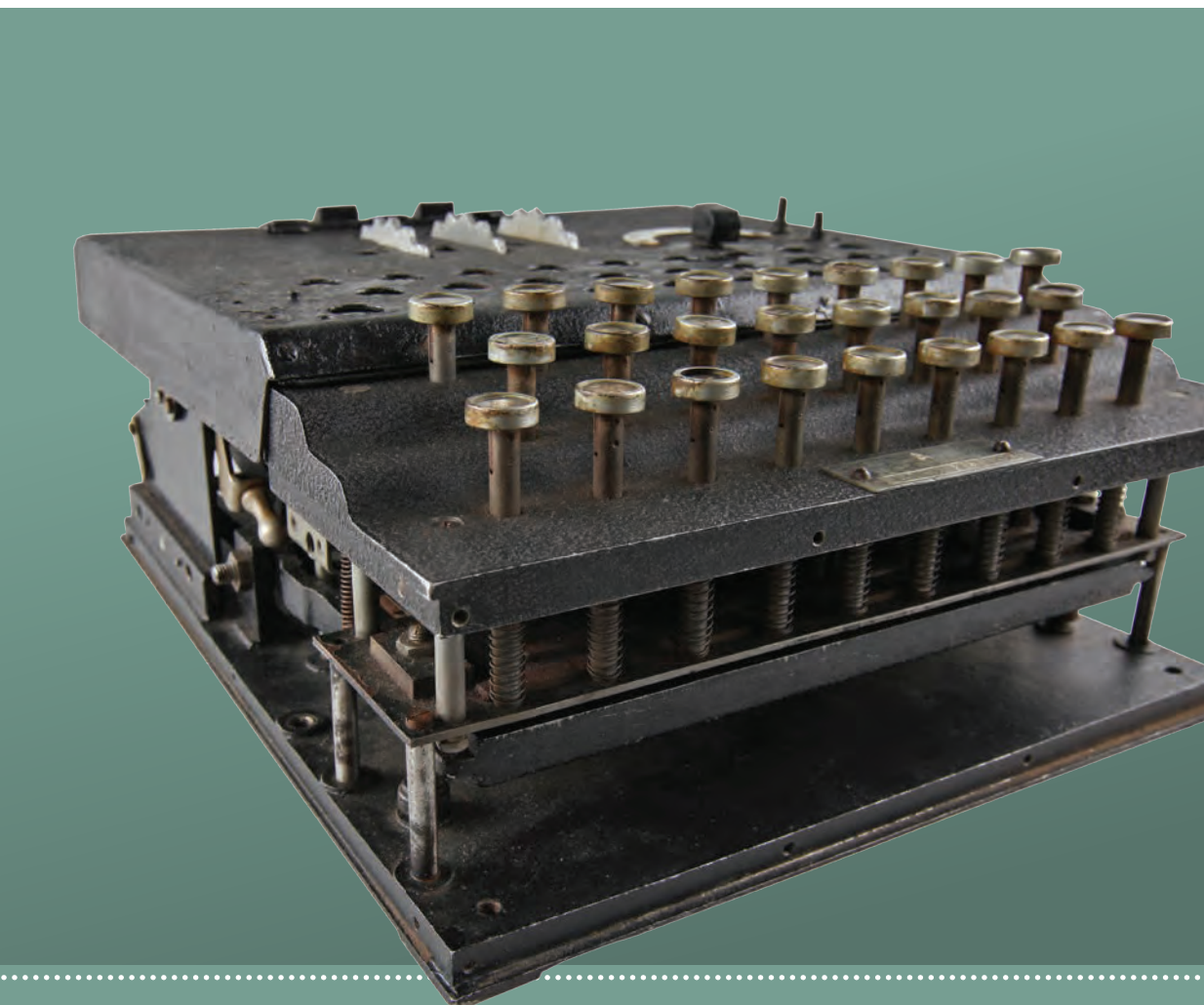
59 | 1 (2019)

PRISPEVKI
ZA NOVEJŠO ZGODOVINO

1



INŠTITUT ZA NOVEJŠO ZGODOVINO

INŠTITUT ZA NOVEJŠO ZGODOVINO

PRISPEVKI ZA NOVEJŠO ZGODOVINO

DIGITAL HUMANITIES AND LANGUAGE TECHNOLOGIES

Letnik LIX Ljubljana 2019 Številka 1

Prispevki za novejšo zgodovino
Contributions to the Contemporary History
Contributions a l'histoire contemporaine
Beiträge zur Zeitgeschichte

UDC
94(497.4) "18/19"
UDK
ISSN 0353-0329

Prevodi/Translations: Studio S.U.R.

Bibliografska obdelava/Bibliographic data processing: Igor Zemljič

Izdajatelj/Published by: Inštitut za novejšo zgodovino/Institute of Contemporary History, Kongresni trg 1, SI-1000 Ljubljana, tel. (386) 01 200 31 20, fax (386) 01 200 31 60, e-mail: jure.gasparic@inz.si

Sofinancer/Financially supported by: Javna agencija za raziskovalno dejavnost Republike Slovenije/ Slovenian Research Agency

Računalniški prelom/Typesetting: Barbara Bogataj Kokalj

Tisk/Printed by: Medium d.o.o.

Cena/Price: 15,00 EUR

Zamenjave/Exchange: Inštitut za novejšo zgodovino/Institute of Contemporary History, Kongresni trg 1, SI-1000 Ljubljana

Prispevki za novejšo zgodovino so indeksirani v/are indexed in: Scopus, ERIH Plus, Historical Abstract, ABC-CLIO, PubMed, CEEOL, Ulrich's Periodicals Directory, EBSCOhost

Številka vpisa v razvid medijev: 720

Fotografija na naslovnici: Enigma, s katero so Nemci med 2. svetovno vojno šifrirali vojaška sporočila, hrani MNZS.

Table of Contents

Editorial

Articles

Reviews and Reports

Editorial Notice

Contributions to Contemporary History is one of the central Slovenian scientific historiographic journals, dedicated to publishing articles from the field of contemporary history (the 19th and 20th century).

It has been published regularly since 1960 by the Institute of Contemporary History, and until 1986 it was entitled Contributions to the History of the Workers' Movement.

The journal is published three times per year in Slovenian and in the following foreign languages: English, German, Serbian, Croatian, Bosnian, Italian, Slovak and Czech. The articles are all published with abstracts in English and Slovenian as well as summaries in English.

The archive of past volumes is available at the **History of Slovenia - SIstory** web portal.

Further information and guidelines for the authors are available at http://ojs.inz.si/index.php/pnz/index.


SISTORY
ZGODOVINA SLOVENIJE

Editorial

Digital Humanities and Language Technologies

The current special issue of the journal *Contributions to contemporary history* brings papers which seem to break with the established editorial tradition. The journal has been issued regularly by the Institute of contemporary history since 1960 which was called *Institute for the History of the Labour Movement* until 1986. The journal was renamed at the same time as the institute, and it has since become one of the major Slovenian scientific journals in the field of history that publishes papers on the contemporary history (19th and 20th century) of Central and Southeastern Europe. With the establishment of an infrastructure programme Research infrastructure of Slovenian Historiography the Institute has entered the field of digital history and has contributed to the establishment of the European *Digital Research Infrastructure for Arts and Humanities* (DARIAH) since 2008. With this, the Institute of contemporary history has started to develop into one of the major digital humanities hubs in Slovenia. The current special issue is one of the results of this new research direction of its publisher and reflects a distinct interdisciplinary and heterogeneous profile of digital humanities.

With this special issue we are celebrating the 20th anniversary of the first *Language technologies conference* which took place in 1998 in Cankarjev dom, Ljubljana and was organized by Tomaž Erjavec, Vojko Gorjanc, Jerneja Žganec Gros and Anica Rant. The topics of the first conference were the development and application of language technologies for Slovene and directions for the future. The conference has since been held biennially and has recently expanded its focus to digital humanities. As the intersection of digital technologies and the humanities, digital humanities is a very active

research field where digital technologies are used in the study of language, society and culture, but humanities research also paves the way for the development of new digital technologies. Digital humanities is a highly interdisciplinary and collaborative field which transforms traditional practices in the humanities and acts as a catalyst of new analytical techniques and methods as well as promotes discussion between the different stakeholders in the field. This initiative aims to promote integration of the disciplines and at the same act as an important hub for fellow researchers in the region.

We invited authors of 11 best-reviewed regular papers and the best student paper that were presented a the *Language technologies and digital humanities conference* which took place on 20–21 September 2018 in Ljubljana, organized by the *Slovenian language technologies society, Centre for language resources and technologies at the University of Ljubljana, Faculty of Electrical Engineering of the University of Ljubljana* and the research infrastructures *CLARIN.SI* and *DARIAH-SI*. Authors of 10 regular papers and the student paper from the fields of language technologies, digital linguistics and digital humanities accepted the invitation and prepared extended papers relevant for an international audience which then underwent another reviewing procedure by international reviewers.

The editors of the special issue would like to thank the authors and the reviewers for their dedicated work as well as for believing in the challenge and being willing to engage in an interdisciplinary dialogue which requires all the parties involved to step out of their comfort zone but also brings knowledge transfer and rewarding results.

Darja Fišer, Andrej Pančur and Tomaž Erjavec
Ljubljana, May 16th 2019

Articles

1.01 UDC: : 004.934:821.163.6-1"16/18"

Nina Ditmajer,[*] Matija Ogrin,[**] Tomaž Erjavec[***]

Encoding Textual Variants of the Early Modern Slovenian Poetic Texts in TEI

IZVLEČEK

ZAPIS VARIANTNOSTI STAREJŠIH SLOVENSKIH PESNIŠKIH BESEDIL V TEI

V prispevku obravnavamo problematiko zapisa verza in variantnih mest v znanstvenokritični izdaji Foglarjevega rokopisa, štajerske baročne pesmarice iz sredine 18. stoletja. Najprej prikažemo diplomatični zapis verza v izbranih problematičnih primerih. V nadaljevanju predstavimo metodo, uporabljeno za izdelavo kritičnega aparata variantnih mest. Temeljno besedilo, tj. Foglarjev rokopis, je primerjano z verzijami v osmih drugih rokopisih in tiskih iz 18. in začetka 19. stoletja. Variantna mesta so označena z elementi XML po Smernicah TEI (TEI Guidelines) kot enote kritičnega aparata. Prikazujemo nekaj primerov detajliranega označevanja rime, stopice, zamenjav verzov ter variantnih razlik na pravopisni, glasoslovni in leksikalni ravnini jezika. Na koncu orišemo več možnosti spletnega prikaza elektronskega diplomatičnega besedila. Pokazala se je potreba po prilagodljivosti teh orodij slovenskemu literarnemu izročilu.

Ključne besede: slovensko slovstvo, Foglarjev rokopis, znanstvenokritična izdaja, kritični aparat, variantnost besedila, TEI

[*] **Research Centre of the Slovenian Academy of Sciences and Arts, Novi trg 2, SI-1000 Ljubljana, nina.ditmajer@zrc-sazu.si**

[**] **Research Centre of the Slovenian Academy of Sciences and Arts, Novi trg 2, SI-1000 Ljubljana, matija.ogrin@zrc-sazu.si**

[***] **Department of Knowledge Technologies, Jožef Stefan Institute, Jamova Cesta 39, SI-1000 Ljubljana, tomaz.erjavec@ijs.si**

ABSTRACT

The paper deals with the problem of encoding the verses and textual variants in the critical edition of Foglar's Manuscript, a Styrian Baroque hymn book from the mid-eighteenth century. We first show the diplomatic transcript of the verse in selected problematic cases, after which we present the method applied to produce a critical apparatus for approaching textual variants. The base text, i.e. Foglar's Manuscript, is compared with versions in eight other manuscripts and prints from the eighteenth and early nineteenth centuries. Variants are encoded with XML elements according to the TEI Guidelines as units of the critical apparatus. We highlight some examples of the detailed encoding of rhymes, feet, verse replacements, and textual variants on the spelling, vocabulary and lexical levels of the language. To conclude, we present a number of possibilities for the online display of the electronic diplomatic transcript. The need for the adaptability of these tools to the Slovenian literary tradition is evident.

Keywords: Slovenian literature, Foglar's Manuscript, critical edition, critical apparatus, textual variance, TEI

Introduction

The texts that have been passed down to us over time via manuscript culture were transcribed from witness to witness over a long period of time. In this kind of textual transmission (*Textüberlieferung*), many textual variations appear in the text, which are called (variant) readings (*Lesarten*) or variants (*Überlieferungsvarianten*). Variant readings can be merely scribal mistakes or "errors", but even these can range from using the wrong letter to the omission of an entire line. Variants, however, can also be the scribe's intentional modifications of the text, including anything from orthographic differences and various word forms to major interventions in the text, such as additions, omissions, word order changes, transpositions of whole paragraphs or stanzas, etc. Textual variance also occurs in printed texts in general, that is, in the culture of the printed book: as soon as the same text is published again, variant readings start to appear, albeit not quite as extensively as in the handwritten tradition. Since very few medieval manuscript texts are preserved in the Slovenian language, the problem of textual variation in Slovenian only appears in the early modern age, especially in the Baroque era. Among the most common examples of the Slovenian transcription tradition are those of the Baroque texts of the eighteenth and nineteenth centuries. Among prose texts, for example, the *Črnovrški Manuscript*,[1] the manuscripts on the *Antikrist*[2] and the *Poljane Manuscript*[3] are mentioned in the present paper, while handwritten

1 The text is treated in the *Register of Baroque and Enlightenment Slovenian Manuscripts* (NRSS Ms 124).
2 Cf. the *Register of Baroque and Enlightenment Slovenian Manuscripts* (NRSS Ms 15, Ms 17, Ms 24, Ms 71).
3 Cf. the *Register of Baroque and Enlightenment Slovenian Manuscripts* (NRSS Ms 23, Ms 28).

hymn books were particularly popular among the common people. These hymn books were preserved through the textual transmission in all of the regional varieties of the Slovenian standard language[4] existing in the Slovenian ethnic territories until the unification of the Slovenian standard language in the mid-nineteenth century. They were either copied by scribes from earlier printed or handwritten hymn books, flyers for special occasions (e.g., pilgrimage, church consecration), lectionaries, catechisms and prayer books, or were written from memory, or dictation.

It is precisely by supplying scholarly evidence and an explanation of its textual tradition that the critical edition should provide us with the most authentic and complete version of a literary work's text: "*When a text is transmitted through more than one witness, a critical edition will generally take a strong interest in recording the variant readings of some or all of those manuscripts or editions*" (Burghart 2017).

Therefore, in addition to the original text, the critical edition should also hand down a textual tradition of witnesses, which exists in the form of transcripts, fragments, drafts, proof sheets, etc. in order to clarify the process of the text's transformation and genesis: "*The apparatus is a set of notes designed to foster in the reader an awareness of the historical and editorial processes that resulted in the text he or she is reading and to give the reader what he or she needs to evaluate the editor's decisions*" (Damon 2017, 202). In principle, digital editions offer more possibilities than printed versions to present the text in its various formats, as they allow for the juxtaposition of different forms of text (for example, a digital facsimile and a diplomatic transcript) in a selected size category and in precisely selected places, at the level of the paragraph, the stanza or the verse (Ogrin 2005, 9–10).

In the present paper, taking as an example the diplomatic transcript of a selected hymnal manuscript, we present the question of encoding the variant readings of the text as reflected in its handwritten and printed versions according to the *TEI Guidelines* from 2019. These can be used to produce a variety of digital texts, from simple reading editions to scholarly critical editions, dictionaries and language corpora. The digital markup means that the structural elements of the text (e.g., verses, stanzas, notes) are encoded with TEI-defined tags that the computer can then recognise. The TEI recommendations consist of descriptions of the tags rendered in the XML markup language, which can be defined as an open encoding standard focused not on the display but on the structure and internal relations of the data. We can use these tags to mark in the electronic encoding the desired structure and other characteristics of the text (Ogrin and Erjavec 2009; Ogrin 2005, 14; Hockey 2000, 24). In this way, we have, since 2004, prepared nine editions of the eZISS library – *Digital Scholarly Editions of Slovenian Literature* (Ogrin and Erjavec 2009).

In the following paragraphs, we present *Foglar's Manuscript*, the selected base text, in a diplomatic transcript, along with its variant readings in the preserved versions of the hymns in other manuscripts and prints. The diplomatic transcript is important not

4 The Eastern Slovenian standard language with its Prekmurje and Eastern Styrian varieties and the Central Slovenian standard language with its Carniolan and Carinthian varieties.

only for locating the original version of the text, but also for comparing versions on all levels of the language. By using suitable web tools, we can also study the stanza forms, verse and metre. In addition to a presentation of selected tools, we were interested in the different kinds of display of the digital diplomatic text in the HTML layout.

The Text Corpus

Foglar's hymn book (1757–1762) is a Slovenian Baroque manuscript containing twenty-four hymns. It originates in the area of the then Austrian province of Styria in the parish of Kamnica near Maribor. The manuscript is named after Lovrenc Foglar, one of its authors (cf. Ditmajer 2017), and contains the following hymn texts: the oldest Slovenian hymns celebrating the pilgrimage to *Mariazell* in Upper Styria; four hymns dedicated to saints; a festive hymn dedicated to the Holy Trinity; two hymns with eschatological content; one worshipping Jesus' name; one of repentance for the fasting period; and another praising the love of God. During the examination of preserved Slovenian religious hymns known to date, as well as other witnesses containing hymn texts, a number of hymns were discovered that could have served as a base text for *Foglar's Manuscript*, or vice versa.

To date, we have included eight variant texts in the critical edition:

- the hymnal manuscript *Pesmarica from Gorje* (1761–1792, NRSS Ms 113),
- Paglovec's hymnal manuscript *Cantilenae variae partim antiquae partim* (1733–1759, NUK R 0 75843),
- Lavrenčič's printed *Misijonske pesme inu molitve* (1757, NUK GS 0 10212),
- Krebs's hymnal manuscript (1750–1800, NRSS Ms 022),
- the hymnal manuscript *Cerkvene pesmi in molitve* (ok 1778, NRSS Ms 052),
- Maurer's hymnal manuscript (1754, NUL Ms 1485),
- Parhamer's printed catechism entitled *Obchinzka knisicza zpitavanya teh pet glavnih stukov maloga katekizmussa* (1764, UKM R 20675), and
- *Manuskript iz Podmelca* (1802–1810, Archives of the ZRC SAZU Institute of Ethnomusicology, Kokošar's Series, Ms. II., Sg. Ms. Ko. 101/125).

The selected variant hymns were mostly produced in the eighteenth century in the regions of Styria, Carinthia, Carniola and Gorizia. Eleven of the hymns exist in a single version (for example, *Pesem od svete trojce, Pesem od božje lubezni, Pesem od svete Notburge*), and only one exists in two versions (*Pesem od Marije Magdalene*). All of the manuscripts and prints mentioned are listed among the listWit (*witness list*) source list added to the preface to the critical edition, and shown as follows:

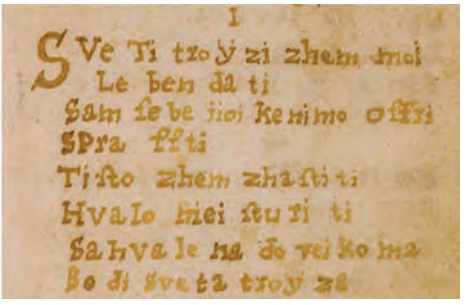
```
<witness xml:id="G">
  <label>Gorje</label>
  <bibl><title xml:lang="sl" type="editorial">Pesmarica iz Gorij</title>,
  <date notBefore="1761" notAfter="1792">1761–1792</date>,
  <idno type="NRSS">Ms 113</idno></bibl>
</witness>
```

The Diplomatic Transcript of the Base Text and Its Variations

In early Slovenian hymn books, one graphic line does not always correspond to a single metric verse. Frequently, due to a lack of paper space, scribes would write the next word or phrase on a second graphic line. In the diplomatic transcript, we used the TEI element `label` to number stanzas; verse lines encoded with an `l` (*line*) are embedded in an `lg` (*line group*) element following `label`; the refrain is nested in the parent stanza (i.e., `lg`) with an assigned `@type` attribute; and the break of the verse line is simply marked with an `lb` (*line break*) element, as shown in the encoding example of the first stanza of *Pesmi od Svete trojce*:

```
<label>1</label>
<lg>
  <l>Sve Ti tro ÿ zi zhem moi <lb/>Le ben da ti</l>
  <l>Sam ſe be jioi kenimo offri <lb/>Spra ffti</l>
  <l>Tiſto zhem zha ſti ti</l>
  <l>Hvalo niei ſtu ri ti</l>
  <lg type="refrain">
    <l>Sahva le na do vei ko ma</l>
    <l>Bo di sve ta troÿ za</l>
  </lg>
</lg>
```

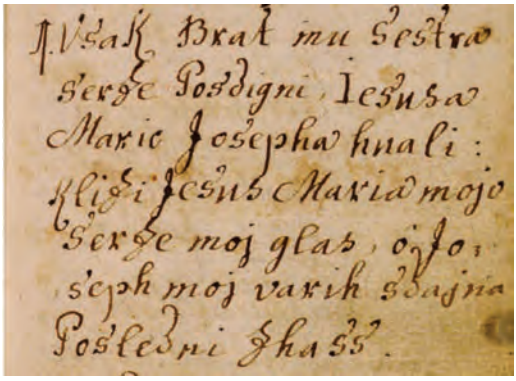
Figure 1: The original variant of the first stanza of *Pesmi od Svete trojce*



Difficulties are caused above all by hymn texts in which the author has disregarded the verse line, rendering the hymn in prose form. In view of this, hymns with a second verse line continuing in the same graphic line where the first verse line begins were encoded with the `ab` (*anonymous block*) element, while the `@type` attribute was used to mark the stanza, with line breaks indicated as shown in the following example:

```
<ab type="lg"><label>1.</label>
  <lb/>Vsak Brat inu Sestra <lb/>Serze Posdigni, Iesusa <lb/>Mario Josepha hvali:
  <lb/>Klizi Jesus Maria mojo <lb/>Serze moj glas, ô Jo-<lb/>seph moj varih sdajna
  <lb/>Posledni zhass.</ab>
```

Figure 2: The original variant of the first stanza of *Pesmi o svetem Jožefu*



In addition to verse lines, stanzas and refrains, rhyme and foot can be specifically encoded in a machine readable format. However, this markup in our scholarly edition have not yet been taken into account. The rhyme patterns can be documented with the `@rhyme` attribute, while the `@label` attribute is used to specify which parts of a rhyme scheme a given set of rhyming words represent. The value of this attribute is usually one of the letters of the rhyme pattern.

```
<lg type="poem" rhyme="aabbcc">
  <l>Sve Ti tro ÿ zi zhem moi <lb/>Le ben <rhyme label="a">da ti</rhyme></l>
  <l>Sam ſe be jioi kenimo offri <lb/><rhyme label="a">Spra ffti</rhyme></l>
  <l>Tiſto zhem <rhyme label="b">zha ſti ti</rhyme></l>
  <l>Hvalo niei <rhyme label="b">ſtu ri ti</rhyme></l>
  <lg type="refrain">
    <l>Sahva le na do <rhyme label="c">vei ko ma</rhyme></l>
    <l>Bo di sve ta <rhyme label="c">troÿ za</rhyme></l>
  </lg>
</lg>
```

In the second example the `@met` attribute indicates the metrical structure, where the symbol | marks the foot boundaries. If some lines divert from the metrical scheme documented in the `@met` attribute, the deviation is documented with the `@real` attribute:

```
<lg met=”u-|u-|u-|u-/”>
  <l>Po ſluſhai kai ti jaſ povem</l>
  <l real=”-u|-u|-u|-u/”>Kai ti ozhem osnani ti</l>
  <l>Nesna nu le tu do vſih mo<unclear>u</unclear></l>
  <l>No tt burgo zhem zha ſti ti</l>
  <l>No tt Burga je Tÿ Rolarza</l>
  <l>S nto lar ſke Do li ne</l>
  <l>Poſhtenih pur garskih ludi</l>
  <l>Prav ſrezhne korenine</l>
    </lg>
```

For a scholarly critical edition of a manuscript, especially one from an early period, it is essential to look for textual variants, as they facilitate the detection of errors in the overall text and aid the search for the base text. In the described critical edition, all of the preserved textual transmissions (traditions) are displayed and organised so as to be subordinate to the base text, that is, Foglar's text. Our first attempt at encoding textual variance in poetic texts was the preparation of the digital critical edition of Anton Martin Slomšek's poems, which was devised in the period 2006–2011 and is still in progress. The diplomatic transcript of Foglar's hymn book was treated with the same *apparatus criticus*, applying the same parallel segmentation method[5] and displaying the variant readings using the `app` element. The latter contains the base text (the lemma), and one or more variant readings encoded with the `rdg` (*reading*) element, each with a reference to the appropriate version via the `@witt` (*witness*) attribute:

```
<l>
  <app>
    <lem wit=”#F”>MAri ia Magda lena</lem>
    <rdg wit=”#M”>An Bart Magdalena</rdg>
    <rdg wit=”#POD”>Enkrat Madalena</rdg>
  </app>
</l>
```

The `@wit` attribute value refers to the identifier of the description of manuscripts and prints with the aforementioned versions of hymnal texts, such as the value "M" for *Maurerjeve pesmarice*, or "POD" denoting the *Manuskript iz Podmelca*, as shown in the

5 For a detailed description of the method, see section "12.2 Linking the Apparatus to the Text" of the TEI Guidelines, *12 Critical Apparatus - The TEI Guidelines*, https://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html.

list of sources in the preceding section. The critical edition includes 988 units of the critical apparatus `app`, which contain 988 `lem` elements and 1072 `rdg` elements. Only pure textual variants were included as units of the critical apparatus, excluding the identification of the verse-stanza structure of the variant text.

Particularly problematic are hymns whose entire stanzas, or simply the verses of a single stanza, are switched, such as in *Pesem od vernih duš*. Such switches can be more explicitly marked using the `@xml:id` (*identifier*) and `@corresp` (*corresponds*) attributes:

```
<l>
  <app>
    <lem wit=”#F” xml:id=”verse5”>Dol vo gen ſo sako pa ne</lem>
    <rdg wit=”#P” corresp=”#verse9”>Vshgala ga je ta praviza</rdg>
  </app>
</l>
<l>
  <app>
    <lem wit=«#F« xml:id=«verse9«>Vuishgalagaje praviza</lem>
    <rdg wit=«#P« corresp=«#verse5«><del>Uſse</del> U' tem ogniu sede
sakopane</rdg>
  </app>
</l>
```

In textual criticism, we distinguish two major groups of variant readings: substantive and accidental (Greg 1950). The latter include those changes that do not significantly affect the meaning, such as orthographic variants, although in some cases even these cause meaning-related dilemmas. The Baroque text of *Foglar's Manuscript* is substantively marked by the non-standard use of spelling and the regional phonetic variation in various branches of the textual transmission. The scope of the critical apparatus and the degree of its granularity have been the subject of discussion in philology since the beginning of critical edition production, especially regarding the distinction between the level of purely orthographic differences, or so-called accidentals, and the level of more meaning-related differences, or so-called substantives, which go back to Greg's theory of copy-text and beyond into the history of philology.[6]

In order to provide a better visual representation of the various types of modification when applying tools for the display and analysis of texts, we need to classify these modifications more precisely and introduce more units of the critical apparatus within one verse line. In the eighteenth century – due to the lack of Slovenian textbooks on spelling and grammar, and of Slovenian books in general, as well as to the fact that school instruction was carried out in a foreign language (only elementary instruction

6 For a comprehensive historical outline of the views that have been formed in textual criticism with regard to this question, see Sahle (2013, 172–73).

was conducted in Slovenian) and that the education of copyists varied – the use of graphic characters for certain sounds varied significantly (marked in the critical edition with the `@type` attribute value):

```
<l>
  <app>
    <lem wit="#F">Bres</lem>
    <rdg wit="#G" type="orthographic">Breſs</rdg>
  </app>
  <app>
    <lem wit="#F">Madesha</lem>
    <rdg wit="#G" type="orthographic">Madeſha</rdg>
  </app>
  <app>
    <lem wit="#F">spozheta</lem>
    <rdg wit="#G" type="orthographic">ſpozheta</rdg>
  </app>
</l>
```

Until the mid-nineteenth century, the Slovenian ethnic territories were characterised by the coexistence of regional varieties of the Slovenian standard language. We therefore encounter many phonological and morphological variant readings in this critical edition, which, like spelling variants, do not affect the meaning of a particular word.

```
<app>
  <lem wit="#F">vun</lem>
  <rdg wit="#P" type="vocalic">ven</rdg>
</app>
```

Lexical substitutions are of more importance, but in the manuscript texts included in the critical edition it is generally a case of synonyms:

```
<l>
  <app>
    <lem wit="#F">dela fairont</lem>
    <rdg wit="#P" type="lexical">Della pust</rdg>
  </app>
</l>
```

Tools for Text Analysis and Display

The XML-TEI encoding of textual variation shown above conveys the logical and semantic structure of the variant readings in the hymns, on the basis of which the editor of the critical edition is able to formulate his or her textological and philological analysis of the textual tradition of a given hymn in a machine readable format. However, this format is not intended for the reading public of the digital edition, that is, for actual reading from the screen. For this purpose, it has to be converted into a reader-friendly display format, such as HTML, where the meaning structure of the text is converted into the appropriate graphic design of the text.

To show textual variance in the textual transmission of *Foglar's Manuscript*, we used (or tested) three tools that have very different sets of functionalities for converting XML-TEI elements to the HTML format of display, and that are derived from very different concepts of the graphic representation of textual variants. Apart from these, *Versioning Machine* (VM)[7] is the tool that probably has the longest history. Although it boasts plentiful functionalities, we did not opt for it in this case because we would have had to extensively adapt the XML format in order for the VM to display it well. The tools were evaluated according to how the relevant files, prepared in strict agreement with the *TEI Guidelines*, were converted without special adjustments.

XSLT Conversion

During the preparation of the digital scholarly edition of Foglar's hymn book, XSLT conversion was predominantly used, having been developed as a working tool for the emerging critical edition of the poems by A. M. Slomšek. A web-based tool[8] supporting this conversion enables the conversion of documents from Word (.docx) into TEI and/or the conversion of TEI documents into HTML. For each conversion, a folder is created that is accessible online and contains both the source file and its converted TEI encoding, as well as the HTML file generated from it. The conversion works so that the general conversion of the TEI encoding (provided and continuously developed by the TEI Consortium) into its HTML version is enriched with local changes that the user can activate by selecting the appropriate profile. For our purposes, we developed a ZRC profile that upgrades the general conversion by placing the variant in braces {}, inside which first a lemma, then a variant reading are listed, separated by a vertical slash |. The name of the version referred to by wit/@witness is displayed when a user places a mouse hover over it.

The aforementioned issue of granularity of the critical apparatus, i.e., how detailed the information about individual variant readings should be (based either on words or larger sections), is clearly shown in Figures 3 and 4. First, Figure 3 shows the solution

7 Cf. *Versioning Machine 5.0*, http://v-machine.org/.

8 *DOCX to TEI to HTML conversion*, http://nl.ijs.si/tei/convert/.

where the `lem` element contains the entire verse of Foglar's text, followed by the `rdg` element containing the whole verse from the manuscript by Mihail Paglovec. In this case, the critical apparatus unit contains and defines the entire verse line as a variant reading. Figure 4, on the other hand, shows the same verse lines as Figure 3, but encoded in a way that each word is represented by its own unit, so each element containing a single word from Foglar's text has a corresponding `rdg` element containing a single word from Paglovec's text. Thus, all of the orthographic and substantive variants are likely to be more clearly shown, with the exception of the spaces between the syllables, which, although not so important for the analysis, does make reading somewhat more difficult.

Figure 3: A synoptic presentation of the base text by Foglar and of Paglovec's variant in HTML format (*Pesem od svete Notburge*)

1
{Po ſluſhai kai ti jaſ povem|Poslushei kar ti jest povem}
{Kai ti ozhem osnani ti|Kar ti zhem osnanite}
{Nesna nu le tu do vſih mou|Nasnano le to do ſedei}
{No tt burgo zhem zha ſti ti|Nottburgo zhem zhastite}
{No tt Burga je Tÿ Rolarza|Nottburga je Tyrolarza}
{ S nto lar ſke Do li ne |S' Intolarske dolline}

Figure 4: A synoptic presentation of the base text by Foglar and of Paglovec's variant in HTML format (*Pesem od svete Notburge*)

1
{Po ſluſhai|Poslushei} {kai|kar} ti {jaſ|jest} povem
{Kai|Kar} ti {ozhem|zhem} {osnani ti|osnanite}
{Nesna nu|Nasnano} {le tu |le to} {do vſih mou|do ſedei}
No tt burgo zhem {zha ſti ti|zhastite}
{No tt Burga|Nottburga} je {Tÿ Rolarza|Tyrolarza}
S {|'} {nto lar ſke|Intolarske} {Do li ne |dolline}

This tool, whose generic conversion according to the *TEI Guidelines* has been upgraded with a synoptic display of the critical apparatus in a main text line, is intended for a simple but philologically accurate presentation of textual variance in a digital scholarly edition. Its use is conditioned by the consistent adoption of the parallel segmentation method in TEI. Although not providing the reader with the greatest flexibility of display (for example, the ability to hide or display a specific version of the text), it is a valuable tool because it is available as an online service[9] and can easily be

9 The conversion service operates at the address http://nl.ijs.si/tei/convert/, by selecting the conversion profile ZRC.

installed on any computer, enabling it to be run at any time during the editorial process. It is ideal for displaying texts in which only two or three, perhaps four, versions are compared in each unit of the apparatus, which seems to be entirely appropriate for the actual range of textual variance established in the earlier Slovenian literary tradition.

TEI CAT

The TEI *Critical Apparatus Toolbox* (TEI CAT) is a web service[10] developed by a group led by Marjorie Burghart. It is explicitly intended for critical editors preparing digital scholarly editions with the parallel segmentation method under the *TEI Guidelines*. It therefore serves as a work aid enabling editors to check and visualise meaning components in the course of the preparation of their scholarly editions. Many functionalities are provided for this purpose, including those for checking errors and inconsistencies that emerge in the encoding process (Burghart 2016). We will focus on the functionalities that are the most relevant to our textual analyses.

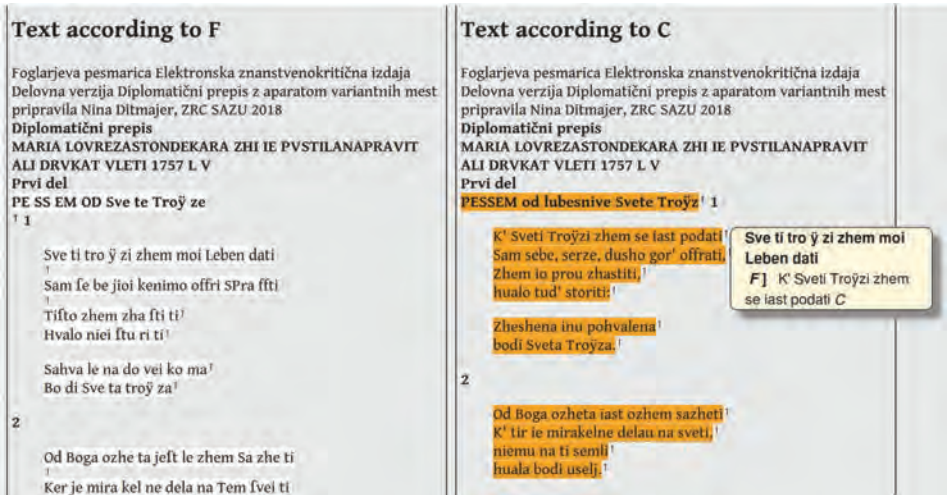
The user sends an XML file to the online service to verify the correctness of the tagging. If the results are positive, the main text or the so-called critical text of the edition will be displayed for viewing. Beside each unit of the critical apparatus, an arrow appears on screen, which can be clicked to open a window with the content of the unit in a classic form based on the use of the right square bracket: everything to the left of the square bracket represents the lemma, while to the right is the variant reading marked with the abbreviation of the variation.

In addition, we are free to select a number of controls, such as whether the system should display page breaks or colour the units of the apparatus that do not contain all of the versions, or, conversely, whether it should colour only those units of the apparatus that contain a specific version, etc.

The most important functionality offered by TEI CAT is a parallel view of all of the versions generated by the tool from the units of the critical apparatus. Regardless of the fact that, according to the *TEI Guidelines*, the recommended place for the list of versions listWit is in the so-called teiHeader metadata element, the CAT system will locate the listWit anywhere in the TEI document (in our case, it is placed in back), logically sorting its information with respect to the abbreviations. The user can then choose to view all of the versions in a parallel display, or, by ticking only the selected abbreviations, have individual versions displayed in parallel for comparison:

10 The consortium developing the tool includes CNRS and the University of Lyon, cf. *TEI Critical Apparatus Toolbox*, http://teicat.huma-num.fr/index.php.

Figure 5: TEI CAT enables the critical editor to view a parallel display of the main text and the selected versions.



The disadvantage of the parallel display in the TEI CAT tool is that, in longer texts, columns match only at the beginning of the file, while in the continuation the relationship can be broken, resulting in the reader losing reference for comparison. The tool cannot (yet) be downloaded to the user's computer and run locally. It is in fact not primarily intended for preparing an edition as a publication for the general readership, but rather serves to allow verifications in the course of the editorial process. However, in addition to its being very practical for displaying the apparatus and several other functionalities, its greatest advantage is the basic statistical analysis that it produces of the document, not just of the TEI tags used, but also of the texts themselves: it generates a simple but informative frequency list of the words occurring in the edition, with any spelling variant being considered as a new word form, of course.

EVT

Open source EVT – *Edition Visualization Technology* – is designed to produce and publish digital scholarly editions in TEI. As with TEI CAT, the encoding of the critical apparatus with the parallel segmentation method is required.[11] A group led by Roberto Rosselli del Turco conceived EVT with the explicit aim of bridging the gap between the *TEI Guidelines* as a first-rate standard for the production of complex philological works, such as critical editions, and the problems that philologists face when they want their editions encoded in TEI visualised and published online (Rosselli del Turco 2014). Whether locally or online, EVT is opened and used as a web page in

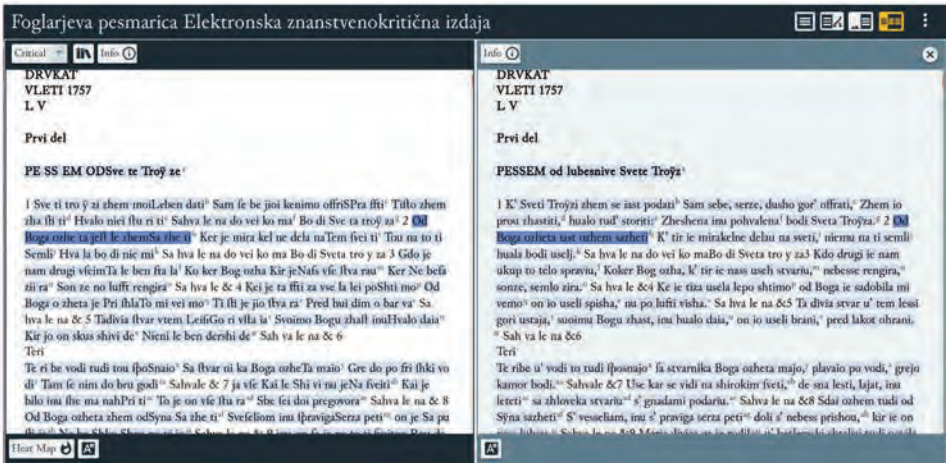
11 The EVT tool is freely available for download to a personal computer and is easy to install.

the selected browser. The tool is designed as a dynamic environment, with Javascript being used to upgrade HTML options. It offers a range of options for displaying critical texts and their variants, including a parallel version and various details about the particular units of the apparatus, which can be freely selected by switching between and generating various displays in real time (see Figure 6). Among the options that would be welcome for the type of editions contained in the eZISS library are support for the dynamic display of digital facsimiles, support for the designated entities and their lists, such as place and personal names, etc. (clearly, these must be appropriately encoded in TEI), and a high level of adaptability to specific project needs.

The conception of the EVT tool is determined by the common conceptual world of Western European philology, whereby the critical editor normally choses to present the text of one selected manuscript accompanied by a smaller or larger number of versions of the same text presented in the form of a critical apparatus. This concept is based on a rich textual tradition composed of thousands of medieval manuscripts both in Latin and in various vernaculars. For example, the digital edition of Chaucer's *Canterbury Tales* prepared by Peter Robinson is based on a transcription of these stories in around 80 preserved manuscripts and incunabula. The Slovenian textual tradition is much less extensive: texts have been preserved in several versions only since the early modern era, while it is only from the eighteenth century onwards that the Slovenian literary tradition offers a significant increase in textual variance. Another large area extremely rich in variation is Slovenian folk poetry, which is not discussed here; nonetheless, EVT might be an ideal tool for studying the exceptional variation of Slovenian folk poetry.

For the Slovenian manuscript culture to which *Foglar's Manuscript* belongs, it is very often the case that only a single manuscript has survived of several witnesses of the text. In such situations, the rich textual tradition has only been passed down to us as one surviving manuscript, the so-called *codex unicus*. This becomes the sole object of a critical edition, which requires a meticulous and detailed presentation, in particular by distinguishing between its diplomatic and critical transcript, which is typical of a philology such as Slovenian philology. In the light of the above, the design of a quality and complex tool, such as EVT, should be appropriately adjusted to optimise the display of a parallel representation of a diplomatic and critical transcript of the same text (in some cases, it will involve critical apparatus, but unless at least two versions of the text have been preserved, the apparatus cannot be compiled).

Figure 6: The EVT tool enables a number of dynamic ways to display the digital scholarly edition, e.g., by showing the main text on the left and the selected version of it on the right.



From this perspective, Foglar's hymn book is a particularly demanding example. On the one hand, with eight previously recorded versions of textual transmission or tradition, it requires a classical Western European type of scholarly edition; on the other hand, a Slovenian philological type of scholarly edition is determined by the contrasting method involving the diplomatic and the critical transcript of the main text. In the future, this need should also be met by adjustments made to its reading display solutions.

Conclusion

The article presents the method adopted to compile a critical apparatus of variant readings in the digital scholarly edition of *Foglar's Manuscript*, a Slovenian Baroque hymn book from the mid-eighteenth century. The editor compared Foglar's text with its versions in eight other manuscripts and old prints. The variant readings identified in the collation process were encoded with XML elements according to the *TEI Guidelines* as units of the critical apparatus. The problem of the variation of older poetic texts raises the problem that various tools embody various functionalities but no tool satisfies the needs of all researchers. This opens up (not entirely new) horizons, where the value of the canonical record of our edition in TEI is further increased, as it can be processed with various, ever evolving tools and according to various needs of presentation and research. Therefore, the first question that arose was how to label a maximum number of analytical findings about the variants using the TEI markup: how to indicate whether the differences are on the level of spelling, vocabulary, lexis, semantics,

etc. The second question was how to best display variants of such diversity in the HTML format designed for reading from the screen. Taking into account the requirements of this critical edition, we tested and evaluated three tools for visualising the critical apparatus. In addition to technology-related differences and the diverse functionalities of these tools, their dependence on individual philological and manuscript traditions has also been shown. As well as the critical apparatus of variant readings, the Slovenian handwritten tradition requires support for the parallel presentation of a diplomatic transcript (with the apparatus) and a critical transcript intended for the wider reading public due to the significant orthographic differences between early and modern Slovenian. In further work we will continue to attempt to further bring the Slovene text tradition ever closer to an ideal method of displaying and publishing texts.

Sources and Literature

- Burghart, Marjorie. 2016. "The TEI Critical Apparatus Toolbox: Empowering Textual Scholars through Display, Control, and Comparison Features." *Journal of the Text Encoding Initiative* 10 (2016). https://journals.openedition.org/jtei/1520#article-1520.
- Burghart, Marjorie. 2017. "Textual Variants." In *Digital Editing of Medieval Texts: A Textbook.* Edited by Marjorie Burghart.
- "Online course: Digital Scholarly Editions: Manuscripts, Texts, and TEI Encoding - Digital Editing of Medieval Manuscripts." *Digital Editing of Medieval Manuscripts*. https://www.digitalmanuscripts.eu/digital-editing-of-medieval-texts-a-textbook/.
- Cankar, Izidor. 2007. *S poti. Elektronska znanstvenokritična izdaja.* Edited by Matija Ogrin, Luka Vidmar and Tomaž Erjavec. Elektronske znanstvenokritične izdaje slovenskega slovstva [Scholarly Digital Editions of Slovenian Literature], ZRC SAZU, IJS. http://nl.ijs.si/e-zrc/izidor/.
- Damon, Cynthia. 2016. "Beyond Variants: Some Digital Desiderata for the Critical Apparatus of Ancient Greek and Latin Texts." In *Digital Scholarly Editing: Theories and Practices*, edited by Matthew James Driscoll and Elena Pierazzo, 201–18. Cambridge: Open Book Publishers.
- Ditmajer, Nina. 2017. "Romarske pesmi v Foglarjevi pesmarici (1757–1762)." In *Rokopisi slovenskega slovstva od srednjega veka do moderne,* edited by Aleksander Bjelčevič, Marija Ogrin and Urška Perenič, 75–82. Ljubljana: Znanstvena založba Filozofske fakultete. http://centerslo.si/wp-content/uploads/2017/10/Obdobja-36_Ditmajer.pdf.
- Ditmajer, Nina, and Matija Ogrin. 2018. "Foglarjeva pesmarica [Foglar's Hymn Book]. Ms 123." In *Register slovenskih rokopisov 17. in 18. stoletja* [Register of Baroque and Enlightenment Slovenian Manuscripts]. http://ezb.ijs.si/nrss/.
- Greg, W. W. 1950. "The Rationale of Copy-Text." *Studies in Bibliography* 3: 19–36.
- Hockey, Susan. 2000. *Electronic Texts in the Humanities*. Oxford: Oxford University Press.
- Ogrin, Matija. 2005. "Uvod. O znanstvenih izdajah in digitalni humanistiki." In *Znanstvene izdaje in elektronski medij*, edited by Matija Ogrin, 7–21. Ljubljana: Založba ZRC, ZRC SAZU.
- Ogrin, Matija, and Tomaž Erjavec. 2009. "Ekdotika in tehnologija: elektronske znanstvenokritične izdaje slovenskega slovstva." *Jezik in slovstvo* 54, No. 6 (2009): 57–72.
- Ogrin, Matija, and Tomaž Erjavec. 2009. "Elektronske znanstvenokritične izdaje slovenskega slovstva eZISS: metode zapisa in izdaje." *Infrastruktura slovenščine in slovenistike*, Simpozij Obdobja 28, edited by Marko Stabej, 123–28. Ljubljana: Znanstvena založba Filozofske fakultete. http://www.centerslo.net/files/file/simpozij/simp28/Erjavec_Ogrin.pdf.
- Ogrin, Matija, and Andrejka Žejn. 2016. "Strojno podprta kolacija slovenskih rokopisnih besedil: variantna mesta v luči računalniških algoritmov in vizualizacij." *Zbornik konference Jezikovne*

tehnologije in digitalna humanistika, edited by Tomaž Erjavec and Darja Fišer, 125–32. Ljubljana: Znanstvena založba Filozofske fakultete, Jožef Stefan Institute.
- "P5 Guidelines – TEI: Text Encoding Initiative." *TEI Consortium*. http://www.tei-c.org/Guidelines/P5/.
- Rosselli Del Turco, Roberto, Giancarlo Buomprisco, Chiara Di Pietro, Julia Kenny, Raffaele Masotti, and Jacopo Pugliese. 2014. "Edition Visualization Technology: A Simple Tool to Visualize TEI-based Digital Editions." *Journal of the Text Encoding Initiative* 8. https://journals.openedition.org/jtei/1077.
- Sahle, Patrick. 2013. *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 1: Das typografische Erbe.* Norderstedt: BoD.
- TEI Consortium. 2018. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 3.3.0. [31 Jan. 2018].

Nina Ditmajer, Matija Ogrin, Tomaž Erjavec

ENCODING TEXTUAL VARIANTS OF THE EARLY MODERN SLOVENIAN POETIC TEXTS IN TEI

SUMMARY

In the process of textual transmission (*Textüberlieferung*), many textual variations appear in the text, which are called (variant) readings (*Lesarten*) or variants (*Überlieferungsvarianten*). The problem of textual variation in Slovenian literary history, which is particularly evident in numerous handwritten and printed hymn books, only appears in the early modern age, especially in the Baroque era. Hymnal texts were transmitted among the people both through oral and written traditions. In the present paper, taking as an example the diplomatic transcript of Foglar's hymn book, we present the question of encoding the variant readings of this hymnal text as reflected in its handwritten and printed versions according to the *TEI Guidelines* from 2019. The TEI recommendations consist of descriptions of the tags rendered in the currently most widely used XML markup language. We present *Foglar's Manuscript*, the selected base text, whose diplomatic transcript contains a critical apparatus of its variant readings located in the other eight preserved hymn books originating in the four historical Slovenian regions. We first highlight examples of the diplomatic transcript of verse lines, differentiating between the graphic and the verse line. Various elements and attributes can be added for the machine analysis of the text, such as an analysis of stanzas and feet. We then present ways of encoding variant readings, using the parallel segmentation method and focusing on verse line switches within stanzas and on substantive and accidental variants. Considering the fact that Slovenian literary texts were significantly marked by the regional varieties of the standard language prior to its unification in the mid-nineteenth century, including by an orthographic heterogeneity, we decided to introduce a number of units of the critical apparatus within a verse line

and assign each variant reading an `@type` attribute value. In the final section, we present three tools for text analysis and display: our own XSLT conversion tool, the TEI *Critical Apparatus Toolbox* and the open source *Edition Visualization Technology* tool. For the critical edition in question, XSLT conversion, which generates a static web site with a visually separate display of the variant readings in a line, turned out to be reasonably appropriate. The TEI CAT tool provides a very useful parallel display of the variants, but is not intended for final publication.

Generally distinguished by powerful functionalities, the EVT tool should be slightly adjusted for the Slovenian textual tradition, in which the diplomatic and critical transcripts of the same text play the major role. Future technological solutions for digital scholarly editions will have to take into account, in particular, the diverse, complex differences in the structure of both transcripts: the diplomatic transcript, for example, with its specific problems is encoded and shown as a paragraph in which several interventions have taken place; the critical transcript, on the other hand, can display the same text in linguistically regularised forms, as a stanza of rhymed verse with a marked metric structure, etc. The parallel representation of the digital facsimile and two methodologically completely different transcriptions (and possibly even a classical critical apparatus) potentially represents a significant technological problem; however, only such an ecdotic (text-critical) conception of the scholarly critical edition can reveal all of the semantic wealth of early modern Slovenian texts.

Nina Ditmajer, Matija Ogrin, Tomaž Erjavec

ZAPIS VARIANTNOSTI STAREJŠIH SLOVENSKIH PESNIŠKIH BESEDIL V TEI

POVZETEK

V procesu rokopisne preoddaje (*Textüberlieferung, Textual transmission*) nastajajo v besedilu številne razlike, ki jih imenujemo variante (*Lesarten, readings*) ali variantna mesta (*Überlieferungsvarianten, variants*). V slovenski literarni zgodovini se problem variantnosti pojavi še posebej v dobi baroka, ta pa je najbolj vidna v številnih rokopisnih in tiskanih pesmaricah, ki so se med ljudstvom širile tako pisno kot ustno.

V prispevku na primeru diplomatičnega prepisa Foglarjeve pesmarice prikazujemo problematiko zapisa variantnih mest istega besedila v preostalih rokopisnih in tiskanih verzijah po Smernicah TEI (TEI Consortium 2019). Priporočila TEI sestavljajo opisne razlage teh oznak, ki so izražene v trenutno najbolj razširjenem računalniškem označevalnem jeziku (*markup language*) XML. Foglarjev rokopis je v naši izdaji prepoznan kot temeljno besedilo (*base text*), ki smo mu v diplomatičnem prepisu dodali kritični aparat variantnih mest, najdenih v osmih drugih pesmaricah iz štirih

slovenskih historičnih pokrajin. Najprej prikazujemo primere diplomatičnega zapisa verza z razlikovanjem med grafično in verzno vrstico. Za strojno analizo besedila lahko zapisu dodajamo različne oznake in atribute, npr. za analizo rime in stopice. Nato z uporabo metode vzporednega segmentiranja variantnih mest (*parallel segmentation method*) prikazujemo primer zapisa variantnih mest. Še posebej se osredotočamo na označevanje zamenjav verzov v kitici ter substancialnih in akcidentalnih variantnih mest. Ker so slovenska besedila pred poenotenjem slovenskega knjižnega jezika precej pokrajinsko obarvana in izkazujejo tudi neenoten pravopis, smo poskusili znotraj enega verza uvesti več enot kritičnega aparata in variante označiti z vrednostjo atributa `@type`. Na koncu smo predstavili in preizkusili tri orodja za prikaz in analizo besedil: našo lastno pretvorbo XSLT, orodje TEI *Critical Apparatus Toolbox* in odprtokodno orodje *Edition Visualization Technology*. Kot razmeroma primerna se je za našo izdajo izkazala pretvorba XSLT, ki izdela statično spletno stran z vizualno ločenim izpisom variantnih mest v vrstici. Orodje TEI CAT omogoča zelo uporaben vzporedni prikaz variantnih mest, vendar ni namenjeno končnemu publiciranju. Orodje EVT bi bilo potrebno ob že razvitih zmogljivih funkcionalnostih nekoliko prilagoditi za slovensko besedilno izročilo, kjer imata največjo vlogo diplomatični in kritični prepis istega besedila. Bodoče tehnološke rešitve elektronskih znanstvenokritičnih izdaj bodo morale upoštevati zlasti raznolike, kompleksne razlike v strukturi obeh prepisov: diplomatični prepis je denimo s svojimi specifičnimi problemi označen in prikazan kot odstavek, v katerega je posegalo več rok ipd.; kritični prepis pa lahko prikazuje isto besedilo v jezikoslovno regulariziranih oblikah, kot kitico rimanih verzov z označeno metrično strukturo itn. Vzporedni prikaz digitalnega faksimila in dveh metodološko povsem različnih prepisov (in eventualno še klasičnega kritičnega aparata) potencialno predstavlja nemajhne tehnološke probleme; vendar šele takšna ekdotična (tekstnokritična) zasnova edicije razpre vse semantično bogastvo starejših slovenskih besedil.

1.01 UDC: 004.934:821.111SHAK(083.41)

Isolde van Dorst*

You, Thou and Thee: A Statistical Analysis of Shakespeare's Use of Pronominal Address Terms

IZVLEČEK

YOU, THOU IN THEE: STATISTIČNA ANALIZA UPORABE IZRAZOV ZAIMKOVNEGA NASLAVLJANJA PRI SHAKESPEARU

Študija se ukvarja z oblikovanjem napovednega modela, namenjenega ugotavljanju, katere jezikovne in nejezikovne značilnosti vplivajo na izbiro zaimkov v Shakespearovih igrah. V angleščini, ki se je uporabljala v Shakespearovem obdobju, je razlikovanje med YOU in THOU, ki je danes arhaično, še obstajalo. Običajno se navaja, da sta ga določala relativni družbeni status ter osebna bližina govorca in naslovljenca. Vendar pa je treba še ugotoviti, ali bo statistično strojno učenje potrdilo to tradicionalno razlago. Proučuje se 23 značilnosti, izbranih z različnih jezikoslovnih področij, kot so pragmatika, sociolingvistika in analiza pogovora. Trije uporabljeni algoritmi – naivni Bayesov klasifikator, odločitveno drevo in metoda podpornih vektorjev – so izbrani kot ilustrativni nabor možnih modelov zaradi njihovih kontrastnih predpostavk in učne pristranskosti. Opravita se dve napovedi, prva o binarnem (you/thou) razlikovanju in druga o trinarnem (you/thou/thee) razlikovanju. Od vseh treh algoritmov daje najboljše rezultate metoda podpornih vektorjev. Po ugotovitvah so značilnosti, ki najbolje napovejo izbiro zaimka, besede iz neposrednega jezikovnega konteksta. Izkazalo se je, da na napoved zaimka vpliva tudi več drugih značilnosti, vključno z imenom govorca in naslovljenca, razliko v statusu ter pozitivnim ali negativnim mnenjem.

Ključne besede: izrazi zaimkovnega naslavljanja, Shakespeare, korpusno jezikoslovje, digitalna humanistika, statistično modeliranje

* **Lancaster University, i.vandorst@lancaster.ac.uk**

ABSTRACT

This study creates a prediction model to identify which linguistic and extra-linguistic features influence pronoun choices in the plays of Shakespeare. In the English of Shakespeare's time, the now-archaic distinction between you and thou persisted, and is usually reported as being determined by relative social status and personal closeness of speaker and addressee. However, it remains to be determined whether statistical machine learning will support this traditional explanation. 23 features are investigated, having been selected from multiple linguistic areas, such as pragmatics, sociolinguistics and conversation analysis. The three algorithms used, Naive Bayes, decision tree and support vector machine, are selected as illustrative of a range of possible models in light of their contrasting assumptions and learning biases. Two predictions are performed, firstly on a binary (you/thou) distinction and then on a trinary (you/thou/thee) distinction. Of the three algorithms, the support vector machine models score best. The features identified as the best predictors of pronoun choice are the words in the direct linguistic context. Several other features are also shown to influence the pronoun prediction, including the names of the speaker and addressee, the status differential, and positive and negative sentiment.

Keywords: pronominal address terms, Shakespeare, corpus linguistics, digital humanities, statistical modelling

Introduction

For several decades much research has been undertaken on the use of *you*, *thou* and *thee* in Shakespeare's works. However, the results so far have yet to arrive at an exact and conclusive answer regarding how these pronouns were used.

This study combines the strengths of multiple research fields in an effort to determine via hitherto unused methods which linguistic and extra-linguistic features influence the choice of second person singular pronoun (*you* versus *thou* or *thee*) in the plays of William Shakespeare. Prior findings in literary and linguistic studies are utilised to find which features could be relevant in this choice, and tools and applications created for corpus linguistics and computer science are exploited to analyse the data in a more exact way than has so far been accomplished. Through these techniques, I hope to identify which features can contribute to a more accurate prediction of pronoun choice, in a model to mimic the pronoun use of Shakespeare.

It is worth observing at this point that it has not yet been determined whether it is even possible to predict the pronoun based on linguistic features. Part of the aim of this paper is to make a determination on this point. In other words, is it possible to create a computational model that can predict which pronoun will be used based on a set of linguistic and extra-linguistic features taken from the text itself and selected on the basis of knowledge that we have of English in the late 1500s and early 1600s? To

accomplish this, all occurrences of *you*, *thou* and *thee* are extracted from Shakespeare's plays, and every instance is manually coded for 23 linguistic and extra-linguistic features, creating data which will serve to ascertain the answer to this primary question. A second question to be addressed is whether some features perform better as predictors of the pronoun choice than others. Thirdly, the issue of whether the use of different algorithms affects the prediction outcomes will be considered.

Throughout this paper, italicised *you*, *thou* and *thee* refer to specific pronoun forms. However, whereas *you* – in Early Modern English as in contemporary English – does not exhibit any formal variation for pronoun case, *thou* is strictly a nominative form with *thee* as its accusative/dative form. *Thou* and *thee* are therefore related inflectional forms of a single pronoun lemma; *you* exists in variation with both. Small capitals are used to indicate the pronoun lemmas, thus: you and thou, where thou includes both *thou* and *thee*. Whenever discussing pronouns in this paper, I am strictly referring to the singular second-person pronouns *you*, *thou* and *thee* that are examined in this study.

Background

Digital Humanities

Over the past few years, computational research has branched out into other research fields that are not necessarily closely connected to computer science. Digital Humanities (DH) is an umbrella term for all research that is computational but approaches the datasets investigated within, and/or addresses questions or problems that are of importance to, the disciplines of the humanities.

The popularity of Digital Humanities, a cross-domain field of study, is attributable to the fact that it does not diminish the differences between fields but rather operationalises this difference to solve difficulties that could not be dealt with within a single discipline. The role of computational methods in the humanities can be considered as that of a supporting character; in any DH computer modelling research, it should be kept in mind that the interpretation is as important at the suitability of a computational model and its outcomes.

Early Modern English and YOU/THOU

In Early Modern English (EModE), two different second person singular pronouns were used, namely the formally singular thou and the formally plural (but pragmatically also respectful-singular) you, with only the latter surviving the EModE period (Taavitsainen and Jucker 2003). The difference between the uses of these two pronouns is evident from multiple literary studies that have addressed Shakespeare's

work, work of his contemporaries, and other documents from this era, such as Walker (2003) and Busse (2002). These studies suggest that unwritten social rules governed the use of these pronouns, abiding by which rules was necessary in order to speak according to society's standards. The use of the two different pronouns acted as a sign of relative status: you would be used to superiors and thou towards inferiors. The choice of pronoun can thus also operate as a subtle means of showing respect or disrespect; using the pronouns in this way would have been natural and easy to English native speakers of the period.

Shakespeare lived during the Early Modern English period, and thus used both you and thou in his writing. His work was written less than 100 years before *thou* and *thee* disappeared from the standard language (surviving in dialects and archaicised registers, such as pious addresses to the divinity). Thus we may straightforwardly posit that the disappearance of thou was likely already in progress around his time. Though obviously heightened in its use of emotional and dramatic language and style to accommodate to the genre of the play script, the language of Shakespeare – including the usage of the two second-person pronouns – can be assumed to be a reasonably good representation of the language used generally in social interaction and conversation at that time (Calvo 1992).

Prior Studies on YOU/THOU

Most studies of Shakespeare's use of YOU and THOU so far have been literary and nonnumeric studies (Brown and Gilman 1960; Quirk 1974; Calvo 1992); the relative few to have used data-based or quantitative techniques did not implement any method beyond directly comparing raw frequency counts (Busse 2003; Mazzon 2003; Stein 2003). Moreover, these studies did not look at all the extant Shakespeare plays, but instead chose a few plays to focus on. Nonetheless, these studies have demonstrated some patterns in the use of YOU and THOU and thus provide a workable foundation for a more in-depth study of the usage of those two pronouns.

These prior studies support in the overall conclusion that the pronouns YOU and THOU appear to be used to support the explicit expression of respect, social status, and familiarity. Quirk (1974) and Mazzon (2003) characterise the role of the pronoun as a linguistic marker, whose usage can be seen as either marked or unmarked. In other words, the use of a particular pronoun can be seen as marked when it is used unexpectedly, for example when YOU is expected based on social status, but THOU is used instead. Thus, in contrast to earlier studies (Brown and Gilman 1960), they do not perceive YOU and THOU to be in direct contrast, and to have a more variable interpretation than was assumed until then, based on the context it occurs in. Calvo (1992) and Stein (2003) expand on this by concluding that markedness of the pronoun is dependent on the context and the situation, in addition to the pronoun choice depending on stable factors such as the social statuses of, and the level of familiarity between, the characters

in Shakespeare's plays; the speakers and addressees in this study – rather than *just* the latter factors (Brown and Gilman 1960). The emotive effect of the utterances within which the YOU/THOU distinction is utilised is of importance as well; feelings such as anger and love for another character may find expression through pronoun choice. This is connected to the notion of respect, as, in an angry remark, marked pronouns can be used to disrespect the addressee based on their social status (Stein 2003).

As Stein (2003) and Busse (2006) already stressed in their studies, a study of YOU and THOU in Shakespeare cannot and should not be limited to a single research discipline. Rather, what is needed is a combination of literature, sociolinguistics, pragmatics and conversation analysis, which are all useful in capturing the complexity of pronominal address and the social constrictions that may have underpinned the choice of one honorific pronoun-form over the other.

Methodology

As has already been mentioned, this is a strictly empirical study which attempts to verify the findings of earlier research through a computational approach. The use of a computational, statistical method is motivated by the goal of creating a more objective representation of Shakespeare's use of YOU and THOU in his plays than has been accomplished so far, since it does not require analysis of meaning-in-context by a human being, but rather proceeds directly from quantitative measurements.

Hypotheses

Three hypotheses were formulated on the basis of the literature:

1. No single model will be able to predict the pronominal address term solely based on linguistic and extra-linguistic features.

This, being a null-hypothesis, is exactly what this study aims to falsify by developing such a model. It is not likely that a single model will be able to predict Shakespeare's original choice of YOU or THOU based on linguistic and extra-linguistic features, because this choice is dependent on so many factors. However, the application of literature, sociolinguistics, pragmatics and conversation analysis all combined into a computational model will be able to successfully predict the pronoun choice as it includes all the factors that might influence the choice for either YOU or THOU.

2. The features of social status, age and sentiment will be better predictors of the pronoun choice than other features.

A hierarchy will be established according to which the linguistic and extra-linguistic features are predicting the pronoun choice in the best performing model. It may be inferred from the literature that social status, age and sentiment are highly likely to be at the top of this hierarchy, among the most influential features; these three features have shown up most reliably in prior research.

3. The best performing algorithm will combine features both dependently and independently.

The different learning biases and assumptions of the three algorithms applied in this study will reveal how the features interact with one another. The first algorithm, Naive Bayes, assumes all features are independent of one another, while the decision tree algorithm assumes that the features are all dependent on each other. Lastly, the support vector machine works with both dependent and independent features. I expect the set of features that will be included in the final model to be a combination of both dependent and independent features, and therefore the support vector machine algorithm to perform best. The three algorithms will be discussed in more detail later in the chapter Classification based on three algorithms.

Data

The data for this study comes from the *Encyclopaedia of Shakespeare's Language* project[1], which is a research project at Lancaster University (UK). The project corpus consists of 38 of Shakespeare's plays, which includes all 36 plays from the First Folio with the addition of *The Two Noble Kinsmen* and *Pericles: Prince of Tyre*. A broadly annotated version of the full Shakespeare corpus can be found online[2]. Some of the annotation and all of the abbreviations used for the titles of the plays follow *The Arden Shakespeare*.

Linguistic and Extra-linguistic Features

The Encyclopaedia of Shakespeare's Language corpus is richly annotated. However, some additional annotation was necessary to perform a full analysis of what extra-linguistic features could be predictors of the pronominal address term. The full set of features used in this study can be found in Table 1. The added features are briefly described here.

As a referent (such as a second person singular pronoun) is dependent on context, the adjacent part of the utterance is used as a feature to test the effect of co-text. Six

1 More information on this project, which is funded by the Arts and Humanities Research Council (AH/N002415/1), can be found on http://wp.lancs.ac.uk/shakespearelang/.

2 CQPweb Main Page, http://cqpweb.lancs.ac.uk.

Table 1: List of all features used in this study

Feature	Acronym	Annotation
Genre	Genre	Pre-annotated
Play name	Play	Pre-annotated
Play, act, scene	Scene	Pre-annotated
Speaker ID	S_ID	Pre-annotated
Speaker gender	S_Gender	Pre-annotated
Speaker status	S_Status	Pre-annotated
Production date	Prod_Date	Pre-annotated
N-gram	LW1-3, RW1-3	Automatic
Positive sentiment	Pos_Sent	Automatic
Negative sentiment	Neg_Sent	Automatic
Speaker age	S_Age	Manual
Location	Location	Manual
Addressee ID	A_ID	Automatic
Addressee gender	A_Gender	Pre-annotated
Addressee status	A_Status	Pre-annotated
Addressee age	A_Age	Manual
Status differential	Stat_Diff	Automatic
No. of people addressed	A_Number	Pre-annotated

co-textual words are included, i.e. a 7-gram altogether. "LW" labels the words occurring on the left of the pronoun, and "RW" the words on the right of the pronoun. Each of these words are numbered based on their distance from the pronoun, e.g. LW3 is the third word on the left of the pronoun. In corpus linguistics, collocations are often examined within a three-word-window, meaning there are three words on either side of the word of interest. While I am not necessarily looking at specific collocations of YOU and THOU, the LW/RW features will look at similarities and differences in co-textual words to see if they can predict the pronoun choice.

Another feature noted as critical in prior studies is sentiment, that is the use of the pronoun to convey positivity or negativity. Sentiment was annotated with the use of the 7-gram described above. *SentiStrength* is a lexicon-based sentiment analysis program that scores phrases with a score for positivity and negativity (Thelwall et al. 2010). Since *SentiStrength* was developed to work with online comments rather than complete sentences as in formal written English, it works well with n-grams too. The scores for positivity and negativity are kept as separate variables.

The corpus already included metadata on the speakers; however, I wanted to include age as well. The age of a character is often not given except for when it is an important attribute of that character, making this difficult to annotate. Therefore, Quennell and Johnson's (2002) character descriptions were used. The characters were

sorted into a trinary classification, with 'adult' as the default category. Any deviations towards 'younger' or 'older' were based on textual references or the character's name, such as for 'Old Man' in *King Lear*. Older characters were occasionally classified as such based on the fact they had adult children with prominent roles in the plays.

A more global feature is the location where the scene is set. This was difficult to annotate, due to the often unreliable stage directions. Instead of a nominal description for each scene location, I used a binary annotation of 'public' and 'private'. The text itself was examined to determine the location based on what characters said about their location, but in addition Bate and Rasmussen's (2007) annotation and Greenblatt, Cohen, Howard and Maus' (1997) annotations were consulted. The use of these three resources enabled the binary manual annotation of location for every scene.

Besides the information about the speaker and the scene, information regarding the addressee is essential when analysing character interaction from a conversation analysis perspective. As a manual annotation for addressee would be incredibly time consuming, I instead used an automatic method which identifies the previous speaker as the addressee of any given utterance. This is in line with the last-as-next bias used in conversation analysis (Mazeland 2003). This means that, even in larger group conversations, it is often expected that the last speaker before the current speaker will also be the next speaker, thus making it likely that the current speaker is addressing the last speaker. If the utterances were interrupted by the start of a new scene or other stage directions (e.g. someone walking into the scene), the annotated addressee would be the *next* speaker rather than the previous speaker for the first utterance after the interruption.

Using the data for the social status of the speaker and the addressee, I also created a status differential. As the status category labels are numeric and ordered, this can be done by taking the difference between the two. For example, a king (status = 0) and a servant (status = 6) are distant in status, and thus will have a high status differential (here: 6). Between a king and a prince (status = 1), the difference is a lot smaller (here: 1). This absolute feature was automatically generated from the already annotated features.

A feature that had to be excluded is familiarity between characters (social distance). This data was not already available, and it was beyond the scope of this study to annotate this for all relevant character pairs. The literature has shown this to be a relevant feature. However, through the use of sentiment analysis, I have attempted to cover the complimentary and insulting aspects that could arise from high familiarity, and any lack thereof arising from low familiarity. Obviously, this does not cover all aspects of familiarity, but it means that this feature is not totally neglected.

Classification Based on Three Algorithms

Three different algorithms are used for the classification task, namely Naive Bayes, decision trees and support vector machines. Whereas it would be ideal to achieve a high precision and recall score, the main goal of this research is to see whether it is even possible to predict the second person singular pronoun choice through a computational application *at all*. If this is indeed the case, what features contribute to this prediction? It is thus more important to verify which features influence the choice and to what extent they do so.

The reason for using three algorithms, and in particular these three, is their differences in learning biases and assumptions. Naive Bayes assumes all features are independent of one another, whereas decision tree attempts to create a dependent, hierarchical structure in the features. Support vector machine (SVM) is more complex and is able to combine both dependent and independent features. The addition of the latter algorithm will be particularly useful if the difference between the two simpler algorithm's models is small.

As well as applying three algorithms, I will also look at the difference between keeping *thou* and *thee* separate and combining them into the one category THOU. For this, I will run both a binary (YOU and THOU) and a trinary (*you*, *thou* and *thee*) classification, to see whether this affects the scores or changes which features are included in the best models.

Overview of Implementation

I ran the three algorithms using the Waikato Environment for Knowledge Analysis (Weka[3]) software[4] with the default settings. The algorithms were run using a 10-fold cross-validation to ensure the best model based on training and testing of all folds combined.

The number of relevant instances of *you/thou/thee* extracted from the dataset is 22,932, which makes up 99.5% of the total number of such pronouns in the dataset. The pronouns were extracted using a Python script with simple heuristics. About 0.5% was missed due to noise in the dataset. The number of instances of *you/thou/thee* that were extracted from each play range from 363 (in *Macbeth*) to 811 (in *Coriolanus*).

I attempted to improve or maintain the scores while making the model simpler by excluding features, that is, through feature ablation. When there were conflicting changes in the scores, the scores of precision and F-measure were prioritised. I hoped to identify which features truly help predict the pronoun by building the simplest but best performing model. The baseline that the models were compared to is derived

3 Weka 3 - Data Mining with Open Source Machine Learning Software in Java, http://www.cs.waikato.ac.nz/ml/weka/.

4 In Weka, Naive Bayes is identified as NaiveBayesMultinominal, decision tree as J48, and support vector machine as SMO.

from the distribution of the pronouns in the dataset, thus 62.6% of YOU and 37.4% THOU.

I first took out groups of features that are related, rather than one feature at a time. Among the 23 features, I created six different groups. The first group related to the wider linguistic and social context (play, production date, genre, scene, location), while the second group was the closer linguistic co-text (n-gram). Information on the speaker (name, status, gender, age) and the addressee (name, status, gender, age, number of people) were groups 3 and 4. I kept status differential on its own, because it relates to multiple groups. Finally, the last group was sentiment (positive and negative). After the group ablation, I went back over the features to see if individual feature exclusions would improve the model further. This ensured the simplest and best model for each algorithm. The scores and the features included in each model are given in Tables 2, 3 and 4.

Results

Trinary Classification Scores

Table 2 shows the results of the trinary classification. As can be seen, each model performed significantly better than the baseline model, on all scores. The F-measure of the best model, the support vector machine model, is highlighted in bold.

Table 2: Scores for precision, recall, F-measure and accuracy for trinary pronoun prediction

Algorithm		Precision	Recall	F-measure	Accuracy
Baseline	Weighted Avg.	0.392	0.626	0.483	62.6417%
	you	0.626	1.000	0.770	
	thou	0.000	0.000	0.000	
	thee	0.000	0.000	0.000	
Naive Bayes	Weighted Avg.	0.826	0.826	0.826	82.64%
	you	0.880	0.885	0.882	
	thou	0.865	0.850	0.857	
	thee	0.509	0.510	0.510	
Decision Tree	Weighted Avg.	0.732	0.752	0.712	75.2093%
	you	0.738	0.960	0.835	
	thou	0.896	0.574	0.700	
	thee	0.408	0.097	0.157	
Support Vector Machine	Weighted Avg.	0.854	0.857	**0.854**	85.675%
	you	0.871	0.927	0.898	
	thou	0.919	0.836	0.876	
	thee	0.659	0.566	0.609	

Binary Classification Scores

Table 3 shows the results of the best models for the binary classification. The F-measure of the best model, again the support vector machine model, is highlighted in bold. This is also the best scoring model out of all models presented in this paper.

Table 3: Scores for precision, recall, F-measure and accuracy for binary pronoun prediction

Algorithm		Precision	Recall	F-measure	Accuracy
Baseline	Weighted Avg.	0.392	0.626	0.483	62.6417%
	YOU	0.626	1.000	0.770	
	THOU	0.000	0.000	0.000	
Naive Bayes	Weighted Avg.	0.868	0.868	0.867	86.8306%
	YOU	0.876	0.920	0.897	
	THOU	0.853	0.782	0.816	
Decision Tree	Weighted Avg.	0.818	0.818	0.818	81.8376%
	YOU	0.849	0.863	0.856	
	THOU	0.764	0.744	0.754	
Support Vector Machine	Weighted Avg.	0.872	0.873	0.872	87.2798%
	YOU	0.886	0.914	0.900	
	THOU	0.848	0.803	0.825	

Feature Comparison of the Models

Overall, the final models contain similar sets of features. The exact compositions are given in Table 4. What is surprising is that the binary classification model for the decision tree is very different from the other models: it does not contain any of the words from the n-gram as a predictor, whereas the others did.

Table 4: Features included in the best model of each algorithm

Algorithm	Type	Features included
Naive Bayes	Trinary	LW1, LW2, RW1, RW2, S_ID
	Binary	LW1, LW2, LW3, RW1, RW2, RW3, A_ID
Decision Tree	Trinary	LW1, LW2, RW1, RW2, S_ID, Stat_Diff, Neg_Sent
	Binary	Scene, S_ID, S_Gender, A_ID, A_Status, A_Age, Stat_Diff, Pos_Sent
Support Vector Machine	Trinary	LW1, RW1, S_ID, S_Age, A_ID, A_Age, A_Number, Stat_Diff, Pos_Sent, Neg_Sent
	Binary	LW1, RW1, S_ID, S_Age, A_ID, A_Age, A_Number, Stat_Diff, Pos_Sent, Neg_Sent

Discussion

This study has given some new insights into the analysis of pronominal address terms. Looking at the second person singular pronoun choice as a binary and a trinary classification problem resulted in slightly different outcomes. Even though the highest scores were achieved in the binary classification, one might still wonder whether this is the best method for addressing the second person singular pronoun choice. Looking back at prior studies on pronoun interpretation and comparing them to the features used in this study, we can conclude that *thee* and *thou* are equal in their opposition to *you*, with the main difference being their grammatical role. From the model comparison, we have seen that the co-text is most important when predicting the pronoun. This is evidence of the purely grammatical difference between *thou* and *thee* and their overall similarity in other aspects. Therefore, both linguistically and computationally, it makes more sense to perform a binary classification.

Differences between the algorithms were observed, but all three algorithms easily outperformed the baseline. The support vector machine models performed best, but the scores for the Naive Bayes models were quite similar to those for the SVM models. A choice between these approaches could be based solely on the scores for accuracy, precision, recall and F-measure, or also by taking into account the complexity, which is significantly higher for the support vector machine models. The more nuanced models that the support vector machine creates, which include more features than the models of the other algorithms, may suggest that the extra complexity of SVM models is indeed beneficial.

The best predicting features were the LW and RW features, which supports the importance of the direct linguistic co-text. In particular RW1 appeared as the most important feature in predicting the second person singular pronominal address term. Other important features were the speaker's name, addressee's name, status differential, positive sentiment and negative sentiment, with additional support from the speaker's gender, addressee's status, addressee's age, speaker's age, and number of people addressed. Only six features were not included in any of the models: genre, play, production date, location, speaker's status and addressee's gender.

I am, therefore, now able to falsify the null-hypothesis that it is not possible to build a reliable prediction model based on linguistic and extra-linguistic features. All six models demonstrate that linguistic and extra-linguistic features substantially improve the prediction of the pronominal address term, as all six outperform the baseline.

The second hypothesis, about which features would be good predictors, was partially correct in predicting that social status, age and sentiment would be included in the best models. However, none of these features were the main predictor of pronoun choice; that was the immediate co-text.

With regard to the final hypothesis, it has been revealed that the features are indeed both dependent on and independent of each other. However, since the Naive Bayes

models perform almost identically to the support vector machine models, we can say that the features are, for the most part, independent of one another.

Conclusions

The primary finding of this study is that it is indeed possible to build a prediction model for the use of YOU versus THOU with a singular referent in the plays of Shakespeare that is based on linguistic and extra-linguistic features. Moreover, in particular, the direct linguistic co-text of the second person singular pronoun is important. Other important features include the speaker's and addressee's names, status differential and both positive and negative sentiment. All in all this suggests that the pronoun choice is influenced by several linguistic and extra-linguistic features.

The best scoring algorithm and model was the support vector machine with 87.3% accuracy through its binary classification model.

For future research, I would recommend an exploration of other algorithms and features that were left out of this study, such as morphology, word embeddings and POS-tags. This will help us gain more information about the linguistic co-text directly surrounding the second person singular pronoun, which will likely give more insight into why this direct co-text is so important in deciding the choice of YOU or THOU. Moreover, including familiarity between characters (social distance) as a feature would be beneficial, as this has been noted multiple times in prior research as an influential factor, but was beyond the scope of this study.

Although this study has not yet provided a comprehensive set of all the linguistic and extra-linguistic features that influence the second person singular pronoun choice in Shakespeare's plays, it has definitely provided a more objective and extensive analysis of the matter that furthers the research into YOU and THOU.

Acknowledgements

The research presented in this article was conducted in collaboration with the Encyclopaedia of Shakespeare's Language project at Lancaster University. This project is funded by the UK's Arts and Humanities Research Council (AHRC), grant reference AH/N002415/1. The Shakespeare corpus will be made publicly available in Summer 2019, first via the CQPweb interface and then through download at a later stage. Many thanks to Jonathan Culpeper and the rest of the team for their advice and support throughout the study.

References

Literature:

- Bate, Jonathan, and Eric Rasmussen, eds. 2007. *William Shakespeare: Complete Works*. London: The Royal Shakespeare Company.
- Brown, Roger W., and Albert Gilman. 1960. "The Pronouns of Power and Solidarity." In *Style in Language*, edited by Thomas A. Sebeok, 253–76. Cambridge: MIT Press.
- Busse, Beatrix. 2006. *Vocative Constructions in the Language of Shakespeare*. Amsterdam: John Benjamins.
- Busse, Ulrich. 2003. "The Co-occurrence of Nominal and Pronominal Address forms in the Shakespeare Corpus: Who Says Thou or You to Whom?", in *Diachronic perspectives on Address Term Systems*, edited by Irma Taavitsainen and Andreas H. Jucker, 193–221. Amsterdam: John Benjamins.
- Busse, Ulrich. 2002. *The Function of Linguistic Variation in the Shakespeare Corpus: A Corpus-based Study of the Morpho-syntactic Variability of the Address Pronouns and Their Socio-historical and Pragmatic Implications*. Amsterdam: John Benjamins.
- Calvo, Clara. 1992. "Pronouns of Address and Social Negotiation in As You Like It." In *Language and Literature*, Vol. 1(1), 5–27. London: Longman Group UK Ltd.
- Greenblatt, Stephen, Walter Cohen, Jean E. Howard, and Katherine E. Maus. 1997. *The Norton Shakespeare: Based on the Oxford Edition*. New York: W.W. Norton & Company, Inc.
- Mazeland, Harrie. 2003. *Inleiding in de conversatieanalyse*. Bussum: Coutinho bv.
- Mazzon, Gabriella. 2003. "Pronouns and Nominal Address in Shakespearean English: A Socio-affective Marking System in Transition." In *Diachronic Perspectives on Address Term Systems*, edited by Irma Taavitsainen and Andreas H. Jucker, 223–49. Amsterdam: John Benjamins.
- Quennell, Peter, and Hamish Johnson. 2002. *Who's Who in Shakespeare*. London: Routledge.
- Quirk, Randolph. 1974. "Shakespeare and the English language." In *The linguist and the English Language*, edited by R. Quirk, 46–64. London: Edward Arnold.
- Stein, Dieter. 2003. "Pronomial Usage in Shakespeare: Between Sociolinguistics and Conversation Analysis." In *Diachronic Perspectives on Address Term Systems*, edited by Irma Taavitsainen and Andreas H. Jucker, 251–307. Amsterdam: John Benjamins.
- Taavitsainen, Irma, and Andreas H. Jucker. 2003. "Introduction." In *Diachronic Perspectives on Address Term Systems*, edited by Irma Taavitsainen and Andreas H. Jucker, 1–25. Amsterdam: John Benjamins.
- Thelwall, Mike, Kevan Buckley, Georgious Paltoglou, Di Cai, and Arvid Kappas. 2010. "Sentiment Strength Detection in Short Informal Text." *Journal of the American Society for Information Science and Technology*, 61(12): 2544–58. https://doi.org/10.1002/asi.21416.
- Walker, Terry. 2003. "*You* and *Thou* in Early Modern English Dialogues: Patterns of usage." In *Diachronic Perspectives on Address Term Systems,* edited by Irma Taavitsainen and Andreas H. Jucker, 309–42. Amsterdam: John Benjamins.

Isolde van Dorst

YOU, THOU AND THEE: A STATISTICAL ANALYSIS OF SHAKESPEARE'S USE OF PRONOMINAL ADDRESS TERMS

SUMMARY

Much research has been undertaken on the use of *you*, *thou* and *thee* in Shakespeare's works. However, the results so far have yet to arrive at an exact and conclusive answer regarding how these pronouns were used. This study combines the strengths of multiple research fields in an effort to determine via hitherto unused computational methods which linguistic and extra-linguistic features influence the second person singular pronoun choices in the plays of Shakespeare. In the English of Shakespeare's time, the now-archaic distinction between YOU and THOU persisted, and is usually reported as being determined by relative social status and personal closeness of speaker and addressee. However, even between studies with similar outcomes, the results vary massively on the degree of influence and by the inclusion or exclusion of a wide range of other potential influencing factors. Therefore, it remains to be determined whether statistical machine learning will support this traditional explanation.

In this study, 23 linguistic and extra-linguistic features are investigated, having been selected from multiple linguistic areas, such as pragmatics, sociolinguistics and conversation analysis. The three algorithms used, Naive Bayes, decision tree and support vector machine, are selected as illustrative of a range of possible models in light of their contrasting assumptions and learning biases. Two predictions are performed, firstly on a binary (YOU/THOU) distinction and then on a trinary (*you*/*thou*/*thee*) distinction, giving six final models to compare. This is a strictly empirical study, which attempts to verify the findings of earlier research through a computational approach. Its aim and main focus is to try and find a pattern or model that best explains the use of second person singular pronominal address terms in Shakespeare, rather than simply achieve the best performing model.

The primary finding of this study is that it is indeed possible to build a prediction model for the use of singular second person pronouns in the plays of Shakespeare based on linguistic and extra-linguistic features. Moreover, in particular, the direct linguistic context of the pronoun is the most important feature in all of the models except one. Several other features are also influencing the pronoun prediction, including the names of the speaker and addressee, the status differential, and positive and negative sentiment. Additionally, all three algorithms easily outperformed the baseline. Out of the three algorithms, the support vector machine models score best. However, the Naive Bayes models perform almost equally well. This reveals that the features are, for the most part, independent of one another. When comparing the binary and trinary classification outcomes, the binary models scored better than the trinary ones.

Looking back at prior studies on pronoun interpretation and comparing them to the features used in this study, we can conclude that *thee* and *thou* are equal in their opposition to *you*, with the main difference being their grammatical role. Therefore, both linguistically and computationally, it makes most sense to use the binary classification.

Isolde van Dorst

YOU, THOU IN THEE: STATISTIČNA ANALIZA UPORABE IZRAZOV ZAIMKOVNEGA NASLAVLJANJA PRI SHAKESPEARU

POVZETEK

O uporabi zaimkov *you*, *thou* in *thee* v Shakespearovih delih je bilo opravljenih veliko raziskav. Vendar rezultati doslej še niso dali natančnega in dokončnega odgovora o tem, kako so se ti zaimki uporabljali. Študija združuje prednosti z različnih raziskovalnih področij, da bi z računalniškimi metodami, ki doslej še niso bile uporabljene, ugotovili, katere jezikovne in nejezikovne značilnosti vplivajo na izbiro osebnega zaimka druge osebe ednine v Shakespearovih igrah. V angleščini, ki se je uporabljala v Shakespearovem obdobju, je razlikovanje med YOU in THOU, ki je danes arhaično, še obstajalo. Običajno se navaja, da sta ga določala relativni družbeni status ter osebna bližina govorca in naslovljenca. Vendar pa se tudi med študijami s podobnimi rezultati ti zelo razlikujejo glede stopnje vplivanja ter upoštevanja ali neupoštevanja številnih drugih mogočih dejavnikov vpliva. Zato je treba še ugotoviti, ali bo statistično strojno učenje potrdilo to tradicionalno razlago.

V tej študiji se proučuje 23 jezikovnih in nejezikovnih značilnosti, izbranih z različnih jezikoslovnih področij, kot so pragmatika, sociolingvistika in analiza pogovora. Trije uporabljeni algoritmi – naivni Bayesov klasifikator, odločitveno drevo in metoda podpornih vektorjev – so izbrani kot ilustrativni nabor možnih modelov zaradi njihovih kontrastnih predpostavk in učne pristranskosti. Opravita se dve napovedi, prva o binarnem (*you/thou*) razlikovanju in druga o trinarnem (*you/thou/thee*) razlikovanju, s čimer dobimo šest končnih modelov, ki jih lahko primerjamo. Študija je strogo empirična, njen cilj pa je z računalniškim pristopom preveriti ugotovitve predhodnih raziskav. Osredotoča se predvsem na iskanje vzorca ali modela, ki bi najbolje pojasnil uporabo izrazov zaimkovnega naslavljanja za drugo osebo ednine pri Shakespearu, in ne le na oblikovanje modela, ki deluje najbolje.

Temeljna ugotovitev te študije je, da je resnično mogoče oblikovati napovedni model za uporabo zaimkov za drugo osebo ednine v Shakespearovih igrah na podlagi jezikovnih in nejezikovnih značilnosti. Poleg tega je neposredni jezikovni kontekst zaimka najpomembnejša značilnost v vseh modelih razen v enem. Na napoved zaimka

vpliva tudi več drugih značilnosti, vključno z imenom govorca in naslovljenca, razliko v statusu ter pozitivnim ali negativnim mnenjem. Vsi trije algoritmi so tudi z lahkoto dosegli boljše rezultate od izhodišča. Od vseh treh algoritmov daje najboljše rezultate metoda podpornih vektorjev. Vendar tudi modeli naivnega Bayesovega klasifikatorja dosegajo skoraj enako dobre rezultate. Iz tega izhaja, da so značilnosti večinoma neodvisne druga od druge. Primerjava binarne in trinarne klasifikacije je pokazala, da so rezultati binarnih modelov boljši od rezultatov trinarnih. Če primerjamo predhodne študije o interpretaciji zaimkov z značilnostmi, uporabljenimi v tej študiji, lahko ugotovimo, da sta zaimka *thee* in *thou* v opoziciji z zaimkom *you* enakovredna, pri čemer je najpomembnejša razlika njihova slovnična vloga. Zato je z jezikoslovnega in računalniškega stališča najbolj smiselna uporaba binarne klasifikacije.

1.01 UDC: 003.295:659.4+004.738.5(497.4))"201"

Darja Fišer,[*] Monika Kalin Golob[**]

Corporate Communication on Twitter in Slovenia: A Corpus Analysis

IZVLEČEK

SLOVENSKO KORPORATIVNO KOMUNICIRANJE NA DRUŽBENEM OMREŽJU TWITTER: KORPUSNA ANALIZA

V prispevku predstavimo korpusno analizo korporativnega komuniciranja na družbenem omrežju Twitter, ki smo jo s kombinacijo besedilnih in metapodatkov izvedli na korpusu Janes-Tweet. Analizirali smo značilnosti slovenskih korporativnih računov in dinamiko njihovih objav ter analizirali rabo novomedijskih elementov in uporabljenega jezika v korporativnih objavah. Na koncu smo proučili še ključne besede v korporativnih objavah. Izvedene analize so pokazale, da v primerjavi z zasebnimi računi v korporativnih tvitih izrazito prevladujejo standardne jezikovne prvine formalnega sporočanja, sicer redkejše neformalne in nestandardne izbire pa so uporabljene premišljeno glede na naslovnika sporočila in namen sporočanja. Prispevek je dragocen tudi zato, ker demonstrira potencial korpusnih pristopov v komunikologiji, medijskih študijah in drugih sorodnih družboslovnih disciplinah, ki proučujejo jezikovno rabo.

Ključne besede: korporativno komuniciranje, družbena omrežja, Twitter, korpusna analiza

[*] **Department of Translation, Faculty of Arts, University of Ljubljana, Aškerčeva 2, SI-1000 Ljubljana, Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, darja.fiser@ff.uni-lj.si**

[**] **Chair of Journalism, Faculty of Social Sciences, University of Ljubljana, Kardeljeva ploščad 5, SI-1000 Ljubljana, monika.kalin-golob@fdv.uni-lj.si**

ABSTRACT

The paper presents a corpus analysis of corporate communication on Twitter, which was performed with a combination of metadata and textual data on the Janes-Tweet corpus. We compare the amount, posting dynamics and use of social-media specific communication elements by Slovene corporate and private users. Next, we analyse the language of corporate users. Our analysis shows that, in comparison to private accounts, corporate tweets predominantly use formal communication and standard language characteristics with seldom usage of informal and non-standard choices. In the event of those, however, they are chosen deliberately to address a specific target audience and meet the desired communicative goals. A major contribution of the paper is also a showcase of corpus-based approaches in communication studies, media studies and other related disciplines in social sciences which study language use.

Keywords: corporate communication, social media, Twitter, corpus analysis

Introduction

In the past decade, social media have evolved into a powerful tool, attracting millions of users every day (boyd and Ellison 2007). Jansen et al. (2010) have shown that around 20 percent of all published tweets mentioned or expressed their opinion about an organization, brand, product or service. What is more, Wu et al. (2011) show that this new form of electronic word-of-mouth is approximately 20 times more effective than marketing events and 30 times more effective than media appearances. It is therefore unsurprising to see such a rapid growth of the online social media marketing (Griffiths and McLean 2014) through which companies address a wide range of goals, such as increased traffic and brand awareness, improved search engine rankings or increased sales (Thoring 2011). In addition, social media can also be used for customer service and market research (Weber 2009).

With the growing commercial relevance of social media, researchers have begun to study the nature and influence of corporate communication on social media. Researchers who investigate the patterns of how information spreads through the Twitter network showed that tweets which contain URLs tend to spread faster (Park et al. 2012) and that tweets containing words which indicate either positive or negative sentiment tend to receive more retweets than neutral posts (Stieglitz and Dang-Xuan 2012). Stelzner (2010) and Heaps (2009) showed that marketers use social media mainly for generating exposure for their business and increasing traffic to their corporate websites, rather than for selling products and services. Evidence has also been found that social media have a positive effect on increasing relational outcomes, such as online reputation and relationship strength (Clark and Melancon 2013; Li et al. 2013; Miller and Tucker 2013). It is therefore surprising that while the new

platform of engagement with customers has shifted the company–customer discourse, Mangold and Faulds (2009) show that communication is still predominantly scripted, promotion-centric and lacks real interaction with the customers.

In this paper we present the results of the first large-scale analysis of corporate communication on Twitter in Slovenia. We look into the production, dynamics and language in the tweets of Slovene corporate users in order to identify the characteristics of such communication in contrast to the communication of private Twitter users. In our study, we use the term corporate account for all private companies, public institutions, the media and interest associations who do not post as individuals for leisure purposes. The analysis was performed on the corpus Janes-Tweet (Erjavec et al. 2018) by combining the available user and text metadata with the content of the tweets, which enabled a more accurate contextualization, parametrization, comparison and generalizations of language use in a specific communicative context.

The rest of the paper is structured as follows: in Section 2 we present related work relevant for our study, in section 3, we present the results of the corpus analysis and Section 4 concludes the paper and outlines future work.

Related Work

In communication studies, three main strands of research into corporate social media communication practices can be identified. The first group focuses on investigating posting behaviour, the second looks into content analysis, and the third are perception studies. In terms of research focus, investigators are mostly interested in corporate communication styles, reputation management and corporate social responsibility.

Quantitative differences in communication dynamics, style and content of Slovene private and corporate Twitter users have been identified by Ljubešić and Fišer (2016) and have been attributed to the different communication functions of private and corporate social media users. While corporate users mostly tweet during the work week in the morning, private users are more active during weekends and in the evening. Corporate tweets have distinctly positive sentiment, while private tweets are predominantly neutral. Tweets posted by corporate users are retweeted much more often while private tweets are more frequently favourited.

By analyzing tweet frequency, following behavior, hyperlinks, hashtags, mentions and retweets, several studies have shown that one-way communication is still the most common communication strategy used by organizations on Twitter (Waters and Jamal 2011; Xifra and Grau 2010) and that the style and genre in tweets by PR professionals is the same as in other PR text types, treating social media as yet another channel for reaching a different consumer segment, without adapting their language accordingly (Kalin Golob et al. 2018). However, as shown by Kwon and Sung (2011), the growing frequency of imperative verb phrases, such as "follow the brand," "come by the booth,"

"join us at the event," or "sign up" for a planned occasion, suggest that corporations increasingly use Twitter as a tool to initiate and maintain relationships with consumers. Risius and Beck (2015) empirically identified social media activities in terms of social media management strategies (using social media management tools or the web-frontend client), account types (broadcasting or receiving information), and communicative approaches (conversational or disseminative). They found positive effects of social media management tools, broadcasting accounts, and conversational communication on public perception. Company account characteristics that have been found to influence public perception are verification, friends, and status.

Gomez and Chalmeta (2013) used content analysis to look into corporate social responsibility (CSR) on social media and have identified presentation, content, and interactivity as the key resources for CSR communication on social media. Presentation refers to the different tools and basic information that supports the company's CSR presence on social media. Content includes messages related to CSR and other topics that reinforce the communication of CSR practices. Interactivity refers to the type of CSR communication and the frequency of CSR messages and feedback.

Li et al. (2013) used social identity theory to identify design factors that determine the social context of a corporate Twitter channel and users' social identification with the community. They confirm that user engagement and informedness in a corporate Twitter channel have a positive effect on corporate reputation and that the credibility of the corporate Twitter channel has a positive effect on user informedness about the corporation. An interesting finding is that deeper relationships among users of a corporate Twitter channel result in higher user engagement and informedness when the level of corporate involvement with the channel is high and the channel has a specific purpose but that the opposite is true when the channel has a generic purpose.

In the related work, post harvesting is typically tailor-made and small-scale, either focused on a few carefully selected corporate social media accounts (e.g. 3 companies), or limited to a carefully designed time span (e.g. 1 month). Coding of the observed phenomena is manual. The research framework is quantitative but done on a relatively small scale, and experimental in that research hypotheses are confirmed or rejected with statistical tests. Our work differs from this research framework in that we use an existing large corpus of tweets and are interested in the characteristics of all the available corporate accounts in it. While coding of certain phenomena (e.g. account type, user gender) was manual, it was performed prior to this study by coders unrelated to this study, so could not be fully controlled. Coding of many other phenomena (e.g. language, sentiment and standardness level of tweets) was automatic and therefore contains a certain degree of noise. Our approach is not only quantitative but large scale as well, taking into account several thousands of users and several million of their tweets, and is descriptive in nature. What is more, unlike most related work which mostly observe the metadata (e.g. tweet frequency, following behavior, retweets) or content of the messages (e.g. hyperlinks, hashtags, mentions, sentiment), we also perform an analysis of the language used in the messages, which is still underresearched

in communication studies. A better understanding of the language practices used by public companies and institutions for presentation, persuasion and reputation management on social media will contribute towards a comprehensive understanding of contemporary, technology-enhanced corporate public relations and marketing strategies and practices. Finally, while most researchers focus almost exlusively on English, our study is performed on Slovene which can serve as a showcase for other languages with a smaller number of speakers (and therefore a smaller market size the corporate accounts are serving).

Corpus Analysis of Corporate Communication on Twitter

The analysis has been performed on the Janes-Tweet corpus (Erjavec et al. 2018) consisting of 11.3 million Slovene tweets or 160 million tokens published by more than 10,200 users. Depending on their communication purpose, users in the corpus are manually divided into two groups: private and corporate. Corporate accounts comprise all private companies, public institutions, the media and interest associations who do not post as individuals for leisure purposes, who are treated as private accounts. In order to establish the characteristics of corporate communication on Twitter and differentiate them from the common practices typical of this medium in general, we perform a contrastive analysis of these two types of accounts.

Our study consists of three parts, each of which addresses a major segment of communication styles on Twitter, ranging from the analysis of communication dynamics and metadata to the content and language analysis, observed from the perspective of the two types of accounts. First, we analyzed the production and posting dynamics of these two user groups. Next, we analyzed the use of social media-specific communication elements, such as hashtags, emojis and emoticons. Finally, we analyzed the language and keywords used in corporate tweets. All the analyses were performed in the SketchEngine corpus-analysis[1] suite (Killgarriff et al. 2014).

The research questions we address with each part of our study are: 1) Does corporate communication on Twitter by Slovene users have a distinct corporate profile in terms of posting dynamics and volume? 2) Have Slovene corporate users adopted the new media communication style and are using the features offered by the new media to maximize their reach and relationship strength? 3) Can we identify the Slovene corporate tweeting code?

1 The corpus is publically available for download as well as for on-line querying through the CLARIN.SI research infrastructure.

Account Analysis

Table 1: Share of corporate and private users and their production in the Janes-Tweet corpus.

Users	No. of users (%)	No. of tokens (%)	No. of tweets (%)
Corporate	2612 (25.57%)	30,003,182 (18.70%)	2,112,910 (18.64%)
Private	7627 (74.44%)	130,401,083 (81.30%)	9,223,736 (81.36%)
Total	10,248 (10.00%)	160,404,265 (100.00%)	11,336,646 (100.00%)

Share of users. The ratio between private to corporate users in the corpus is 3:1. As can be seen in Table 1, less than a fifth of all the tweets in the corpus have been posted by corporate users. This means that in Slovenia, Twitter is mainly used for private communication.

Table 2: Distribution of tweets by corporate and private users based on gender in the Janes-Tweet corpus.

	Corporate		Private	
Gender	**No. of tweets**	**%**	**No. of tweets**	**%**
Unknown	1,730,258	81.89%	134,048	1.45%
Male	271,729	12.86%	6,136,470	66.53%
Female	110,923	5.25%	2,953,218	32.02%
Total	2,112,910	100.00%	9,223,736	100.00%

Users' gender. As shown in Table 2, gender could not be determined for the majority of corporate users (82%) based on user name, user profile data and verb form usage in their tweets, which is rare in the case of private users (1.5%). This is unsurprising because corporate users tweet on behalf of their company or organization, adapting their style of writing accordingly, e.g. the use of first person plural verb forms, which do not distinguish the gender of the writer.

Posting Analysis

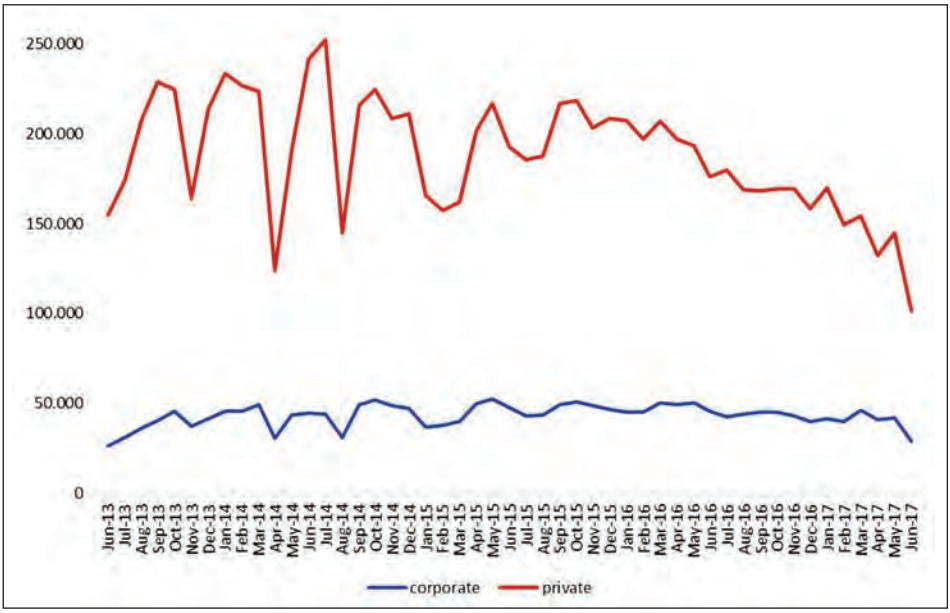
Post quantity. There are only 29 (1%) corporate users who are very active on social media and have posted over 10,000 tweets, and 422 (16%) medium-active ones with 1,000 – 10,000 tweets. The majority of corporate users (1,640 or 62.79%) fall into the category of low-activity accounts with 100 – 1,000 tweets. The lowest-activity group includes 521 users (19.95%) who have posted fewer than 100 tweets. In comparison to private users, the biggest difference is in groups 2 and 4. There are 9% more private users with 1,000 – 10,000 tweets and a similar percentage fewer private

accounts with only 100 – 1,000 tweets. In the years included in the Janes-Tweet corpus, the volume of content generated by the corporate users is stable but is decreasing slightly among the private users (see Figure 1). Occasional sharp drops in the number of posts, which are simultaneous for both user groups, were caused by the technical issues during data collection and are not related to the seasonal fluctuations or other content-related phenomena.

Table 3: Activity of corporate and private users in the Janes-Tweet corpus.

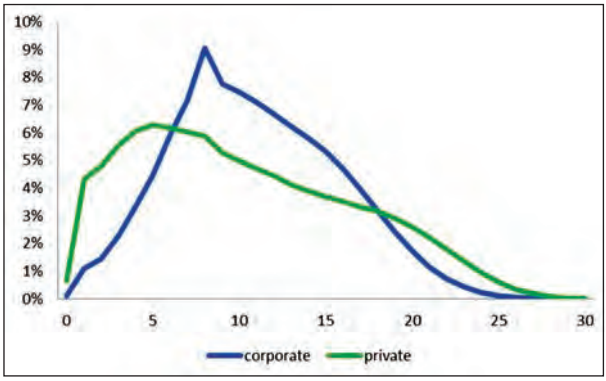
	Corporate		Private	
No. of all accounts	2612	%	7627	%
> 10,000 tweets	29	1.11%	129	1.69%
Between 10,000 and 1,000 tweets	422	16.16%	1867	24.48%
Between 1,000 and 100 tweets	1640	62.79%	4055	53.17%
< 100 tweets	521	19.95%	1576	20.66%

Figure 1: Posting dynamics of corporate and private users in the Janes-Tweet corpus. according to the number of posted tweets between June 2013 and June 2017.



Post length. Figure 2 shows that the length of corporate tweets is more homogenous than the length of private tweets. The biggest share of corporate tweets are 7 to 11 words long (4 to 7 words in case of private users). The share of corporate tweets which do not contain any word (only emojis, hashtags, hyperlinks or multimedia elements) is only 0.1%. Such tweets are six times more frequently produced by private users, which is not surprising as these symbols are typically used in bidirectional communication, which is rare in corporate PR tweets.

Figure 2: Tweet length of corporate and private users in the Janes-Tweet corpus.



Analysis of Interactive Elements

Likes. As can be seen from Table 4, nearly 80% of corporate tweets do not receive any likes, 12% have one like and only 9% have 2 or more likes. Private tweets receive significantly different attention: a third of all the private tweets is liked at least once and a significant share of them (0.7%) receives over 10 likes. This is another strong sign that bidirectional communication is less typical of corporate users and that corporate tweets are just one of the channels of the same type of (one-directional) communication disseminated through different genres.

Table 4: Share of liked and retweeted tweets of corporate and private users in the Janes-Tweet corpus.

No. of likes				
	Corporate users		**Private users**	
	No. of tweets	%	No. of tweets	%
0	1,663,755	78.74%	610,9048	66.23%
1	265,385	12.56%	1,890,549	20.50%
2–10	175,788	8.32%	1,160,057	12.58%
>10	7,982	0.38%	64,082	0.69%
Total	2,112,910	100.00%	9,223,736	100.00%
No. of retweets				
	Corporate users		**Private users**	
	No. of tweets	%	No. of tweets	%
0	1,754,988	83.06%	8,414,713	91.23%
1	219,698	10.40%	490,346	5.32%
2–10	134,184	6.35%	300,319	3.26%
>10	4,040	0.19%	18,358	0.19%
Total	2,112,910	100.00%	9,223,736	100.00%

Figures 3 and 4: The most liked (left) and the most retweeted (right) tweet posted by corporate users in the Janes-Tweet corpus.

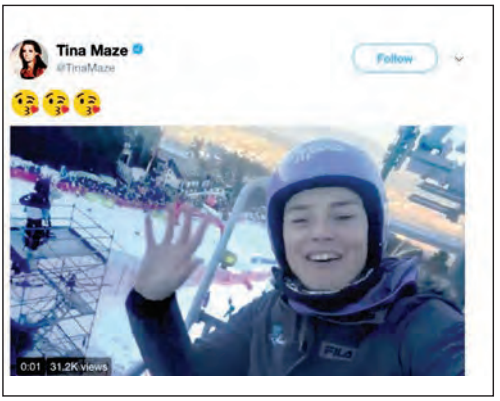


Table 5: Use of hashtags, emoji, hyperlinks and mentions by corporate and private users in the Janes-Tweet corpus.

Hashtags			
	Abs. freq.	Per million	Per tweet
Corporate	922,504	30,746.9	0.44
Private	2,241,693	17,190.8	0.24
Emoji			
	Abs. freq.	Per million	Per tweet
Corporate	1,285,696	42,852.0	0.61
Private	12,061,885	92,498.3	1.31
Hyperlinks			
	Abs. freq.	Per million	Per tweet
Corporate	1,989,643	66,314.4	0.94
Private	2,583,651	19,813.1	0.28
Mentions			
	Abs. freq.	Per million	Per tweet
Corporate	659,211	21,971.4	0.31
Private	9,216,857	57,460.2	1.00

Retweets. Retweeting results show a different picture where a much greater share of corporate tweets have at least one retweet (17%) in comparison to private tweets (8%), suggesting a higher informative value of corporate tweets for a wider audience. Interestingly, when considering very frequently retweeted posts, no difference between the two account types has been observed.

Use of hashtags. Relatively speaking, corporate accounts use hashtags almost twice as often as private accounts. On average, almost every second corporate tweet contains a hashtag, which holds for only every fourth private tweet. As presented in Table 5, sport is the predominant topic of the 10 most frequent hashtags used by corporate users which is very similar to private users. Interestingly, half of the 10 most frequently used hashtags are shared (sport, news, Ljubljana). Among the 10 corporate users with the highest relative frequency of hashtag use we can find less formal magazines and companies. Therefore, for a more detailed analysis of corporate communication it would be interesting to further divide corporate users into different groups: media (journals and magazines), companies, state institutions and non-governmental organizations. We plan to include this in our future studies.

Use of emoticons and emojis.[2] The usage of emoticons and emojis is opposite to hashtags, as emojis are, relatively speaking, more than twice as common in posts by private users who use 1.3 emojis or emoticons per tweet on average while occurring only in every second corporate tweet which indicates greater degree of formality in corporate communication on Twitter. Among the 10 corporate accounts the relative frequency of emojis and emoticons, we mainly identified resellers of fashion items.

As presented in Table 6, all of the most frequently used emojis or emoticons are positive which again indicates a positive tone in PR communication. However, it is interesting that only 2 emojis appear on the top 10 list for corporate users while the rest are emoticons. This could be a sign of more conservative communication strategies used by corporate users given that emojis are a much more recent phenomenon, but this could also be a consequence of corporate users more frequently tweeting from their computers rather than smart phones which better support the use of emojis.

Table 6: Ten most frequent hashtags in corporate and private tweets.

Corporate users		**Private users**	
Hashtag	Frequency	Hashtag	Frequency
#plts	18,03	**#plts**	26,370
#slonews	18,247	**#slonews**	18,270
#PLTS	9,620	**#junaki**	18,167
#Ljubljana	5,724	#slochi	13,195
#izvršba	5,167	**#PLTS**	10,943
#NKDomzale	4,437	#Slovenia	10,780
#olimpija	4,176	**#Ljubljana**	10,141
#rokomet	4,143	#radiobattleSI	9,184
#junaki	3,941	#ligaprvakov	9,091
#skupajdovrha	3,864	#sp14si	8,351

2 Emoticons (e.g. ;)) are combinations of standard typographical characters used for expressing emotions. Emojis are pictograms (e.g. ) which include emotions as well as a broad range of other topics and their usage and interpretation depend on the individual.

Table 7: Ten most frequent emoticons and emojis in corporate and private tweets and the ten corporate accounts with the highest relative frequency of emoticons and emojis.

Emoji	Frequency	User	Frequency	Rel. freq*
:)	114,602	RecycleMan	530	12.711,5
;)	55,763	JennParisBags	188	11.522,1
:D	17,715	EtiVelikonja	160	10.409,8
<3	13,688	ApartmaNet	184	10.104,9
:-)	9,672	TRENDtrgovina	436	10.049,3
;-)	4,926	Pawla40	228	9.720,0
:))	4,680	iPlacesi	125	8.860,0
❤	3,679	bozicluka	92	8.290,2
:P	3,558	matejgaber22	99	7.222,6
☺	3,436	Modniovitki	424	7.010,9

* Relative frequency is the average frequency of the phenomenon in one million tokens.

Table 8: Ten most frequently mentioned accounts in the tweets posted by corporate and by private users.

Corporate users		Private users	
Mention	Frequency	Mention	Frequency
@YouTube	8,325	@petrasovdat	91,328
@Nova24TV	6,903	**@YouTube**	71,859
@Val202	3,992	@MarkoSket	57,333
@rtvslo	3,866	**@JJansaSDS**	53,482
@kzssi	3,736	@lucijausaj	51,391
@unionolimpija	3,616	@leaathenatabako	44,453
@JJansaSDS	3,464	@petrajansa	44,102
@radioPrvi	3,128	@savicdomen	43,394
@vladaRS	2,764	@darkob	42,363
@nkmaribor	2,758	@zzTurk	40,534

Use of hyperlinks. Great differences between private and corporate users can be observed in their use of hyperlinks in tweets. Relatively speaking, corporate tweets contain more than three times the number of hyperlinks in comparison to private tweets. On average, corporate users add a hyperlink to nearly each tweet they post, while private users include it only in every fourth tweet. This corresponds to the findings of our preliminary analysis that tweets are often only compressed press releases leading to a complete message in the form of a hyperlink.

Mentions of other users. Big differences between private and corporate users are observed in the rate and type of other user accounts mentions. Relatively speaking,

mentions are more than twice as frequent in private tweets as they are in corporate tweets. On average, private users mention other users in every tweet, whereas corporate users use this option only in every third message. This is not surprising because the main objective of PR tweets is self-presentation, which is why referencing others is less needed. Among the 10 most frequently mentioned accounts in corporate tweets are mainly media, political institutions/parties/individual politicians and sport organizations, while in private tweets we find social media influencers, two journalists and a politician. Both lists have only two mentions in common, i.e. YouTube and Janez Janša, one of the oldest and best known Slovenian politicians.

Language Analysis

Language of tweets. Corporate users almost exclusively post messages in Slovene (93%), which is considerably different from private users whose share of tweets in a foreign language is twice as large. Among the foreign languages used in tweets of corporate users, English prevails (5%). This corresponds to our preliminary findings that the main goal of Slovene corporate Twitter users is to address their Slovene audience through formal communication for business or informative purposes. The only exception are the accounts of Slovene Embassies around the world often posting in their local language (e.g. in French), as well as the accounts of the Ministry of Foreign Affairs, the president and the prime minister who occasionally use English tweets to inform the international community about major events (e.g. arbitration).

Table 9: Language use in the tweets posted by corporate and private users.

	Corporate		Private	
Language	No. of tweets	%	No. of tweets	%
Slovene	1,973,677	93.41%	8,074,681	87.54%
English	104,955	4.97%	983,141	10.66%
Bosnian/Croatian/Serbian	16,058	0.76%	57,017	0.62%
Other	18,220	0.86%	108,897	1.18%
Total	2,112,910	100.00%	9,223,736	100.00%

Sentiment of tweets. Every tweet in the corpus is annotated with a sentiment label (see Erjavec et al. 2018). Half of all corporate tweets have positive sentiment, a third has neutral sentiment and 17% of the tweets have negative sentiment. This greatly differs from private tweets, half of which are neutral, 27% negative and only a quarter positive. This is another indication of the PR nature of corporate tweets which try to convey a positive corporate image, attract customers, sell products, etc.

Table 10: Sentiment of tweets posted by corporate and private users.

	Corporate		Private	
sentiment	No. of tweets	%	No. of tweets	%
positive	1,024,238	48.48%	2,320,841	25.16%
neutral	729,811	34.54%	4,411,516	47.83%
negative	358,861	16.98%	2,491,379	27.01%
total	2,112,910	100.00%	9,223,736	100.00%

Table 11: Language standardness level in the tweets posted by corporate and private users.

	Corporate		Private	
Standardness	No. of tweets	%	Sentiment	No. of tweets
L1	1,688,244	79.90%	4,515,310	48.95%
L2	353,397	16.73%	3,489,743	37.83%
L3	71,269	3.37%	1,218,683	13.21%
	2,112,910	100.00%	9,223,736	100.00%

Table 12: Comparison of the language used in corporate and private tweets according to part of speech.

Part of speech	Corporate (per million)	Private (per million)	Ratio**
Proper nouns	66,738.4	33,507.8	1.99
Numerals	30,564.9	16,109.7	1.90
Conjunctions	54,381.1	33,302.1	1.63
Prepositions	86,947.2	54,549.6	1.59
Adjectives	76,889.9	48,254.8	1.59
Common nouns	186,446.6	127,056.0	1.47
Abbreviations	3,826.0	3,458.9	1.11
Punctuation	143,234.6	158,188.2	0.91
Main verbs	62,631.9	75,795.7	0.83
Auxiliary verbs	36,974.7	52,968.0	0.70
Adverbs	38,192.1	55,483.1	0.69
Pronouns	39,118.2	62,678.8	0.62
Particles	19,816.6	35,540.7	0.56
Interjections	1,740.9	6,194.5	0.28

** Ratio between the frequency in corporate and in private tweets.

Language standardness. Tweets by corporate users mainly contain standard Slovene (80%) and highly nonstandard content is only rarely present (3%). Almost the opposite is true of private users. Less than half of their tweets are written in standard Slovene and the share of tweets containing highly nonstandard Slovene is more

than four times greater in comparison to corporate users. Some exceptions can be found among the accounts of public personalities (e.g. stand-up comics, radio presenters, musicians) who often purposefully tweet in nonstandard Slovene because informal communication is a major part of their corporate image.

Orthography. Great differences are detected regarding the use of abbreviations: corporate tweets mainly contain standard abbreviations of academic or other titles (*dr., mag., d. o. o.*) and common abbreviations (*št., oz., min.*), while in private tweets we find nonstandard abbreviations (*tw*), often without full stop (*slo, lj, min*). Some differences can be also observed in the use of punctuation. In corporate accounts, a bigger range of classic punctuation marks is used according to the orthographic norm. Tweets by private users are characterized by frequent repetitions of the same punctuation mark to give the message an emotional charge. Much more frequent is also the use of social-media specific symbols (*#, @, **).

Parts of speech. The analysis of the parts of speech in the language of corporate tweets offers an insight into communication purposes of corporate accounts. Relatively speaking, there are almost twice as many proper nouns and numerals in corporate tweets than in private ones. Much more frequent are also conjunctions, prepositions, adjectives and common nouns. As shown in Table 10, interjections are considerably more often present in private accounts (3.5 times more). The same is true for particles (almost 2 times more), pronouns and adverbs. On the one hand this confirms a greater formality of corporate users and reflects a more direct and personal approach of private users. On the other hand this also reflects different communicative functions of Twitter: informative for corporate and conversational for private accounts. Furthermore, the informative, as well as the influencing function to some extent, are also confirmed by the detailed analysis of individual parts of speech presented below.

The noun. Common nouns are 1.5 times more common in corporate tweets than in private ones, but the matching rate of the first 20 common nouns that are most frequently used is surprisingly high (70%): *dan/day, leto/year, tekma/race, ura/hour, mesto/place, teden/week, čas/time, hvala/thank you, svet/world, delo/work, človek/human, konec/end, otrok/child, država/country*. Among the 20 most frequent nouns, the following are specific to corporate tweets: *video/video, foto/photo, zmaga/victory, novica/news, cena/price, sezona/season*. Proper nouns are twice as common in corporate tweets than in the private ones and the matching rate of the 20 most frequent nouns is 40%: *Slovenija/Slovenia, Ljubljana, Maribor, EU, Slovenc/Slovene, Evropa/Europe, ZDA/USA, Cerar, Janša*. Among the 20 most frequent nouns, the following proper nouns are corporate tweets: *Olimpija, Koper, Peter, Gorica, Janez, Domžale, Luka, Tina, Marko*.

In corporate tweets a higher level of formality of expression has been detected as both first and last names are indicated (private tweets mention only the last name). Furthermore, we can observe greater diversity of places and company names. An analysis of nominal pronouns returned predictable results: corporate tweets contain plural pronouns (*nam/to us, nas/us, vam/to you*), while in private tweets we find singular

forms of pronouns (*jaz/I, me/of me, ti/to you, te/you*). The reason for grammatical plurality lies in the fact that authors of corporate tweets use formal communication methods on behalf of their institution or company and formal form of addressing.

The verb. The use of main verbs is more common in private tweets. The matching rate of the 20 most frequent verbs in private and corporate tweets is 60% (*imeti/have, iti/go, morati/must, vedeti/know, videte/see, priti/come, dobiti/get, začeti/begin, čakati/wait, dati/give, praviti/say, delati/work, dobiti/get*), but the difference lies in their motivation for communication: corporate accounts mainly report on events and publish statements, while private accounts describe personal activities and give opinions. Among the 20 most frequent verbs, the following main verbs are specific to corporate tweets: *želeti/wish, preveriti/check, najti/find, iskati/search, prebrati/read, gledati/watch, moči/able, hoteti/want, narediti/do*.

The adjective. Adjectives are 1.5 times more frequently used in corporate than in private tweets and the matching rate of the 20 most used adjectives is 50%: *nov/new, dober/good, slovenski/Slovenian, velik/big, lep/beautiful, zadnji/last, mlad/young, star/old, pravi/real, super/super*. Among the 20 most frequent adjectives the following are specific to corporate tweets: *vabljen/invited, današnji/today's, evropski/European, javen/public, spleten/web/based, svetoven/world/wide, odličen/excellent, državen/national, visok/high, domač/domestic*. Positive adjectives are characteristic of corporate tweets (*nov/new, dober/good, velik/big, lep/beautiful*) which are also more formal than the adjectives characteristic of private tweets (*vabljen/invited, odličen/excellent, visok/high* vs. *hud/badass, mali/little, sam/alone*). Adjectival as well as nominal pronouns are used in the first person plural form in corporate tweets (*naše/our-Female, naši/our-Male*) when the goal is identification with the company or the institution and integration into the communicative circle that connects the author of the message on behalf of the institution with the recipient (Korošec 1998).

The particle. The difference between formality and informality can also be observed through particles which overlap in 80% of the cases. However, among the particles that are present only in tweets of one user group, our analysis showed that formal particles are distinctive for corporate tweets (*morda/maybe, predvsem/above all, sicer/though, skoraj/nearly*) and nonstandard and informal particles for private tweets (*tud < tudi/also; ze < že/already, itak/off course, pač/well*).

The interjection. As already mentioned, the analysis of this part of speech showed most notable differences. The matching rate of the 20 most common interjections in corporate and private tweets is 55%: *bravo, hm, haha, uf, o, ej, ah, ha, aha, aja, oh*. Among the most frequent interjections that are distinctive for one of the user groups are the following ones: *živjo, zdravo, hej, hehe, gooool, opa, ups, na, ojoj*. Interjections in corporate tweets are fewer in quantity as well as more formal and salutatory (*zdravo, ups*), while private tweets often contain interjections in foreign language (*btw, lol*) and swear words.

Keyword Analysis

This section highlights the results of the keyword analysis performed on corporate tweets. In this paper, the keywords are understood as the words which are unexpectedly more frequent in the tweets of corporate users compared to the entire JanesTweet corpus as reference.

Table 13: List of 20 most key lemmas in corporate tweets according to sentiment.

Negative	Keyness index	Positive	Keyness index	Neutral	Keyness index
oviran	22.2	čestitka	3.5	novice.si	10.1
trčenje	19.1	vabljen	3.5	zemljišče	8.7
trčiti	18.0	bravo	3.4	pivniški	8.3
priključek	15.4	album	3.4	ebel	8.3
evakuirati	15.3	beautiful	3.4	katarinin	8.1
ranjen	15.1	hvala	3.4	petv	8.0
poškodovan	15.0	posted	3.4	šloganje	7.9
razcep	14.9	photos	3.4	solaten	7.8
novicejutro.si	14.9	odličen	3.3	ugnati	7.8
osumljen	14.6	polepšati	3.3	pripravljalen	7.7
nesreča	14.5	odlično	3.3	koel	7.6
aretirati	14.3	prijeten	3.3	novinec	7.6
avtocesta	14.1	super	3.3	napovednik	7.4
neurje	14.1	čudovit	3.3	zoofa	7.3
strmoglaviti	13.9	čestitati	3.3	prerokovanje	7.3
osumljenec	13.1	srečno	3.3	poiesis	7.2
magnituda	13.1	facebook	3.3	apod	7.1
prometen	12.8	welcome	3.3	wt	7.1
ubit	12.8	summer	3.3	sklepen	6.9

Sentiment. As shown in Table 13, the highest keyness index is attributed to lexis from corporate tweets with negative sentiment. Among those, all 20 top-ranking key lemmas are part of media tweets that reference reports on crime and other accidents (e.g., *trčenje/collision, evakuirati/evacuate, ranjen/injured, nesreča/accident*). The 20 top-ranking keywords with positive sentiment correspond to the definitions of positive PR communication (e.g., *čestitka/congratulations, vabljen/invited, bravo/bravo, čudovit/wonderful, polepšati/make sbd's (day)*). Adjectives and adverbs with highly positive meaning are also ranked high (e.g., *lep/beautiful, odličen, odlično/fantastic, prijeten/nice, super/super*). Furthermore, the 20 top-ranking keywords with neutral sentiment are part of the tweets containing media reports (e.g., *novice.si/news.si, zemljišče/property, napovednik/preview, sklepen/final*) and denote events (e.g., *pivniški/beer, ebel/ebel, šloganje/card-reading, prerokovanje/fortune-telling*) or names (*katarinin, ebel, zoofa, apod*).

This list suggests that for a more fine-grained analysis of corporate communication on Twitter it could be useful to consider separating the tweets generated by media from those that are created by companies or institutions.

Table 14: Comparison of key word forms in corporate tweets, written in standard and non-standard language.

Standard tweets	Keyness index	Non-standard tweets	Keyness index
Izkl	6.4	Posetite	562.3
Novice.SI	6.4	potrazi	557.6
dražba	6.0	sjajan	553.5
[hyperlink]	5.9	Jeste	455.0
SiOL	5.8	tim	308.5
Petv	5.8	[hyperlink]	307.2
APOD	5.8	[hyperlink]	186.6
Moia	5.7	li	166.4
spletnem	5.7	koketo	145.9
Zurnal24	5.7	trombeto	143.3
ugodne	5.7	[hyperlink]	130.0
astronomska	5.7	belooranžnega	129.5
SMUČANJE	5.6	deejaytime	111.2
KOŠARKA	5.6	Živjo	111.0
oviran	5.6	Skupne	109.6
[hyperlink]	5.6	pritisne	92.8
ALPSKO	5.6	oglasiš	66.2
HOKEJ	5.6	[hyperlink]	65.9
zamudite	5.6	cheers	60.3
Preverite	5.5	hajskul	56.5
Nogometaši	5.5	[hyperlink]	49.6
TENIS	5.5	gnargnar	49.6
ciganskih	5.4	sporočimo	47.0
NOGOMET	5.4	najbrš	46.8
ROKOMET	5.4	pridte	45.3
[hyperlink]	5.4	javimo	41.9
Astrolife.si	5.4	Poslali	41.5
Izbrane	5.4	dm	41.2
Slovenske	5.4	javiš	41.2
SMUČARSKI	5.4	unc	41.0

Standardness. A comparison of the 30 top-ranking key word forms (see Table 14) in corporate tweets written in standard and nonstandard Slovene shows that users write in standard Slovene when posting notifications and adds (e.g., *dražba/auction,*

ugodne/good, zamudite/miss, preverite/check). Tweets written in nonstandard Slovene have a similar communication purpose, but numerous elements in foreign language and nonstandard spelling of Slovene words indicate that authors of such messages want to establish a closer connection with their target audience and make their offer more appealing to them (e.g. *deejaytime/phoneticized spelling of DJ/time, hajskul – phoneticized spelling of high school, najbrš – nonstandard for I guess, pridte –nonstandard for come, dm – abbreviation for direct message, javiš – nonstandard for answer*).

Tabela 15: Comparison of key word forms in corporate tweets written by male and female users.

Female	Keyness index	Male	Keyness index
foodwalks	7.7	Moia	41.7
Posodobljen	7.0	dražba	39.9
Patsy	6.1	APOD	37.2
KOEL	5.9	astronomska	36.4
[hyperlink]	5.9	premičnin	35.4
info@patsy.si	5.5	UGANKA	33.9
[hyperlink]	5.5	[hyperlink]	30.7
foodwalk	5.5	Izhodišče	30.3
Lylo	5.3	FOTOGRAFIJE	30.0
ORTO	5.1	GLASBA	29.6
UriKuri	4.6	Dopolni	29.5
yummy	4.6	UE	29.1
Ordered	4.4	javna	27.5
Shellac	4.4	sedežna	27.2
Cosmo	4.2	GCC	26.5
LPG	3.8	PRIPOROČAMO	26.4
Starševski	3.7	Espargaro	26.4
e-trgovine	3.5	[hyperlink]	26.3
[hyperlink]	3.5	zemljišča	26.0
Elle	3.3	[hyperlink]	25.3
info@tjasaseme.si	3.3	Pomurskem	24.8
boxa	3.2	ENERGIJE	24.5
derivatov	3.2	Žurnal24	24.4
IBU	3.1	LITERATURA	24.3
Onaplus	3.1	gozda	24.2
Aquafresh	3.0	[hyperlink]	23.5
naftnih	3.0	PRS	23.1
Watercolour	3.0	Ekipa24	22.8
[hyperlink]	3.0	[hyperlink]	22.3
foodwalks	7.7	Moia	41.7

Gender. While a comparison of the key word forms from female or male corporate accounts in Table 15 does not offer any insights into possible linguistic differences between them, it does give us information about differences in topics and style in regard to language choices made when addressing female or male target audience. Female accounts include names of magazines, URLs and proper names related to fashion, shopping, food and parenting, while in male account these elements are related to real estate, sport and music.

Conclusions

Social media have revolutionized corporate communications by allowing companies to communicate directly and instantly with their stakeholders, marking a shift from the traditional one-way output of corporate communications, to an expanded dialogue between company and consumer (Matthews 2010). This paper presents the results of the first comprehensive, large-scale and corpus-driven analysis of the characteristics of corporate communication on Twitter in Slovenia that could serve as a starting-point of further, data-driven and linguistically enhanced investigations of the importance of social media for fostering corporate communication. In the study, we combined the analysis of the available metadata, Tweet content and corpus annotations to study three key aspects of the communication of Slovene corporate Twitter users: (1) the participation, posting dynamics and posting volume, (2) the utilization of new media elements, and (3) the language choices observed through several levels of linguistic discription.

Based on the Janes-Tweet corpus, Twitter appears to be mainly used for private communication in Slovenia. The majority of corporate accounts belong to the low-activity category but the volume of content generated by the corporate users is stable. Corporate tweets are more homogenous length-wise and are predominantly longer than those of private users.

The analysis of the usage of the new media elements suggests that corporate tweets come short of the true dialogic approach as most Slovene companies and institutions use Twitter as yet another channel for unidirectional communication of regular (shortened) PR messages, while the prevalent communication function remains informative and positively presentational. This can be seen from a much less frequent usage of emoticons and all other interactive elements typical of private accounts, which display a distinct conversational communication function that can be seen in their frequent usage of non-standard particles, interjections, punctuation and language, and a large number of favourites.

A very strong feature of corporate communication is the almost exclusive usage of Slovene which is undoubtedly strategic with a clear focus on the Slovene market. While standard language and formal elements do prevail in corporate tweets of Slovene companies and institutions, the infrequent occurrences of informal and non-standard elements seem to be used deliberately and tailored to the specific target

audience, which points towards a growing awareness of adapting the style to the content that is communicated (level of formality, linguistic standardness, discursiveness), target audience (general public – neutral style vs. specific public – variations between neutral and colloquial style) and the organization profile (public institution – neutral style, standard language, companies – visible, colloquial, non-standard features).

Both sentiment- and part-of-speech-based keyword analyses show an interesting landscape of corporate tweets. The usage of evaluative adjectives is prominent throughout this subcorpus, among which superlatives stands out in particular. The negative keywords originate from the coverage of accidents and crimes by the media, and the positive fully correspond with the definition of promotional elements. These results indicate an important difference between the negative reporting-style tweets by the news outlets, and the positive promotional style of companies, public institutions and non-governmental institutions, suggesting the need for a more fine-grained categorization of corporate accounts, which will be refined in our future work. We also plan to focus on analyzing the reception of corporate tweets which contain non-standard language and interactive elements which are more typical of private communication on social media.

An important original contribution of this study is its demonstration of the methodological potential of corpus approaches in communication studies, media studies and related disciplines in social sciences which are based on language data, which is not yet utilized in the Slovene context. Apart from theoretical relevance, the results of this analysis therefore also have practical implications for PR practitioners and organizations in that they reinforce the importance of properly trained PR practitioners who use social media in a dialogic, two-way symmetrical model, understand their role as boundary spanners and the need to seek opportunities to engage in and stimulate dialogue with stakeholders. The results of our study also clearly illustrate to the PR practitioners that social media should not be treated as just another means through which to disseminate the same advertisements and publicity pieces that stakeholders are already receiving through other traditional media channels. According to Matthews (2010), social media offers an opportunity for direct and instant corporate communication as well as an opportunity to get back to the ideal basics of public relations – building and maintaining relationships – and to change some of the negative stereotypes typically associated with the industry.

Acknowledgments

The work described in this paper was funded by the Slovenian Research Agency within the national basic research project "Resources, methods, and tools for the understanding, identification, and classification of various forms of socially unacceptable discourse in the information society" (J7-8280, 2017–2019) and the Slovenian-Flemish bilateral basic research project "Linguistic landscape of hate speech on social media" (N06-0099, 2019–2023).

Sources and Literature

- boyd, danah m., and Nicole B. Ellison. 2007. "Social Network Sites: Definition, History, and Scholarship." *Journal of Computer-Mediated Communication* 13 (1): 210–30. doi:10.1111/j.1083-6101.2007.00393.x.
- Clark, Melissa, and Joanna Melancon. 2013. "The Influence of Social Media Investment on Relational Outcomes: A Relationship Marketing Perspective." *International Journal of Marketing Studies* 5 (4): 132–42. doi:10.5539/ijms.v5n4p132.
- Erjavec, Tomaž, Nikola Ljubešić, and Darja Fišer. 2018. "Korpus slovenskih spletnih uporabniških vsebin Janes." In *Viri, orodja in metode za analizo spletne slovenščine*, edited by Darja Fišer, 16–43. Ljubljana: Znanstvena založba Filozofske fakultete.
- Gomez, Lina M., and Ricardo Chalmeta. 2013. "The Importance of Corporate Social Responsibility Communication in the Age of Social Media." In *16th International Public Relations Research Conference*, 1–16. Amsterdam: Elsevier.
- Griffiths, Marie, and Rachel McLean. 2014. "Unleashing Corporate Communications: Social Media and Conversations With Customers." In *UKAIS International Conference Proceedings 2014*, 1–51. https://aisel.aisnet.org/ukais2014/51.
- Heaps, Darrel. 2009. "Twitter: Analysis of Corporate Reporting Using Social Media." *Corporate Governance Advisor* 17 (6): 18–22.
- Jansen, Bernard J., Mimi Zhang, Kate Sobel, and Abdur Chowdury. 2010. "Twitter power: Tweets as electronic word of mouth." *Journal of the American Society for Information Science and Technology* 60 (11): 2169–88. doi:10.1002/asi.21149.
- Kalin Golob, Monika, Nada Serajnik Sraka, and Dejan Verčič. 2018. *Pisanje za odnose z javnostmi: temeljni žanri*. Ljubljana: Založba FDV.
- Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. "The Sketch Engine: Ten Years On." *Lexicography* 1 (1): 7–36. doi:10.1007/s40607-014-0009-9.
- Korošec, Tomo. 1998. *Stilistika slovenskega poročevalstva.* Ljubljana: ČZD Kmečki glas.
- Kwon, Eun Sook, and Yongjun Sung. 2011. "Follow Me! Global Marketers' Twitter Use." *Journal of Interactive Advertising* 12 (1): 4–16. doi:10.1080/15252019.2011.10722187.
- Li, Ting, Guido Berens, and Maikel de Maertelaere. 2013. "Corporate Twitter Channels: The Impact of Engagement and Informedness on Corporate Reputation." *International Journal of Electronic Commerce* 18 (2): 97–126. doi:10.2753/JEC1086-4415180204.
- Ljubešić, Nikola, and Darja Fišer. 2016. "Slovene Twitter Analytics." In *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*, edited by Darja Fišer and Michael Beißwenger, 39–43. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.
- Mangold, W. Glynn, and David J. Faulds. 2009. "Social Media: The New Hybrid Element of the Promotion Mix." *Business Horizons* 52 (4): 357–65. doi:10.1016/j.bushor.2009.03.002.
- Matthews, Laura. 2010. "Social Media and the Evolution of Corporate Communications." *The Elon Journal of Undergraduate Research in Communications* 1 (1): 17–23.
- Miller, Amalia R., and Catherine Tucker. 2013. "Active Social Media Management: the Case of Health Care." *Information Systems Research* 24 (1): 52–70. doi:10.1287/isre.1120.0466.
- Park, Jaram, Meeyoung Cha, Hoh Kim, and Jaeseung Jeong. 2012. "Managing Bad News in Social Media: A Case Study on Domino's Pizza Crisis." In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media Relations Review,* 409–11.
- Risius, Marten, and Roman Beck. 2015. "Effectiveness of Corporate Social Media Activities in Increasing Relational Outcomes." *Information & Management* 52 (7): 824–39. doi:10.1016/j.im.2015.06.004.
- Stelzner, Michael A. 2010. "Social Media Marketing Industry Report: How Marketers are Using Social Media to Grow Their Businesses." Accessed February 15, 2019. http://www.socialmediaexaminer.com/social-media-marketing-industry-report-2010/.

- Stieglitz, Stefan, and Linh Dang-Xuan. 2012. "Impact and Diffusion of Sentiment in Public Communication on Facebook." In *ECIS 2012 Proceedings.* Accessed February 15, 2019. https://aisel.aisnet.org/ecis2012.
- Thoring, Anne. 2011. "Corporate Tweeting: Analysing the Use of Twitter as a Marketing Tool by UK Trade Publishers." *Publishing Research Quarterly* 27 (2): 141–58. doi:10.1007/s12109-011-9214-7.
- Waters, Richard D., and Jia Y. Jamal. 2011. "Tweet, Tweet, Tweet: A Content Analysis of Nonprofit Organizations' Twitter updates." *Public Relations Review* 37 (3): 321–24. doi:10.1016/j.pubrev.2011.03.002.
- Weber, Larry. 2009. *Marketing on the Social Web: How Digital Customer Communities Build Your Business*. Hoboken, New Jersey: Wiley.
- Wu, Shaomei, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. "Who Says What to Whom on Twitter." In *Proceedings of the WWW'11 Conference*, 705–14. New York: ACM. doi:10.1145/1963405.1963504.
- Xifra, Jordi, and Francesc Grau. 2010. "Nanoblogging PR: The Discourse on Public Relations in Twitter." *Public Relations Review* 36 (2): 171–74. doi:10.1016/j.pubrev.2010.02.005.

Darja Fišer, Monika Kalin Golob

CORPORATE COMMUNICATION ON TWITTER IN SLOVENIA: A CORPUS ANALYSIS

SUMMARY

In the past decade, social media have transformed corporate communications by enabling direct and instant communication with the stakeholders. In communication studies, three main strands of research into corporate communication practices on social media can be identified: posting behaviour, content analysis and perception studies. Investigators are mostly interested in corporate communication styles, reputation management and corporate social responsibility. A better understanding of the language practices used by public companies and institutions for presentation, persuasion and reputation management on social media is still lacking.

This paper addresses this gap with the first comprehensive, large-scale and corpus-driven analysis of the characteristics of corporate communication on Twitter in Slovenia. In the study, we combined the analysis of the available metadata, Tweet content and corpus annotations in the Janes-Tweet corpus to study three key aspects of the communication of Slovene corporate Twitter users: (1) their participation, posting dynamics and posting volume, (2) the use of social-media specific communication elements, and (3) the language choices observed through several levels of linguistic discription.

Our analysis shows that, in comparison to private accounts, corporate tweets predominantly use formal communication and standard language characteristics with

seldom usage of informal and non-standard choices. In the event of those, however, they are chosen deliberately to address a specific target audience and meet the desired communicative goals. The analysis of the utilisation of the new media elements by corporate users clearly show that their tweets come short of the true dialogic approach and that most Slovene companies and institutions use Twitter as yet another channel for unidirectional communication of regular (shortened) PR messages in which the prevalent communication function remains informative and positively presentational. A keyword analysis reveals an important difference between the negative reporting-style tweets by the news outlets, and the positive promotional style of companies, public institutions and non-governmental institutions, suggesting the need for a more fine-grained categorization of corporate accounts, which will be refined in our future work.

Another major contribution of the paper is its demonstration of the methodological potential of corpus approaches in communication studies, media studies and related disciplines in social sciences that are based on language data, which is not yet utilized in the Slovene context. Apart from theoretical relevance, the results of this analysis therefore also have practical implications for the PR community which highlight the importance of properly trained PR practitioners who use social media in a dialogic, symmetrical model, understand their role as boundary spanners and the need to seek opportunities to engage in and stimulate dialogue with their stakeholders.

Darja Fišer, Monika Kalin Golob

SLOVENSKO KORPORATIVNO KOMUNICIRANJE NA DRUŽBENEM OMREŽJU TWITTER: KORPUSNA ANALIZA

POVZETEK

V zadnjem desetletju so z omogočanjem neposrednega in takojšnjega stika z deležniki družbena omrežja močno vplivala tudi na korporativno kominiciranje. V komunikologiji korporativne komunikacijske prakse na družbenih omrežjih raziskujejo z opazovanjem vedenja korporativnih uporabnikov, analizo vsebine in percepcijskimi študijami. Komunikologe zanimajo predvsem slogi poslovnega sporočanja, upravljanje ugleda in družbena odgovornost podjetij, medtem ko še vedno primanjkujejo jezikoslovno usmerjene raziskave, ki bi omogočile boljše razumevanje jezikovnih praks, ki jih podjetja in institucije uporabljajo za predstavljanje svojih izdelkov, vplivanje na potrošnike in odzivanje v kritičnih situacijah.

To vrzel naslavlja pričujoči prispevek, v katerem predstavimo prvo celovito, na obsežnem korpusu zasnovano analizo korporativnega komuniciranja med slovenskimi uporabniki družbenega omrežja Twitter. Izvedli smo jo s kombinacijo besedilnih

podatkov, metapodatkov in korpusnih oznak, ki so na voljo v korpusu Janes-Tviti, pri analizi pa smo se osredotočili na tri vidike korporativnega komuniciranja v slovenskih uporabnikov: (1) njihovo prisotnost, aktivnost, dinamiko in količino objav, (2) rabo novomedijskih komunikacijskih elementov in (3) jezikovne izbire, opazovane na različnih ravneh jezikovnega opisa.

Izvedene analize so pokazale, da v primerjavi z zasebnimi računi v korporativnih tvitih izrazito prevladujejo standardne jezikovne prvine formalnega sporočanja, sicer redkejše neformalne in nestandardne izbire pa so uporabljene premišljeno glede na naslovnika sporočila in namen sporočanja. Analiza izkoriščanja novomedijskih elementov jasno kaže, da komuniciranje slovenskih korporativnih uporabnikov na družbenem omrežju Twitter ne sledi dialoškemu pristopu in da večina slovenskih podjetij in institucij Twitter razume kot dodatni kanal za enosmerno sporočanje klasičnih (skrajšanih) sporočil za javnost, sporočanjska vloga katerih ostaja pretežno informativna in pozitivno predstavitvena. Analiza ključnih besed razkrije pomembno razliko med negativnim poročanjskim slogom medijskih računov in med pozitivnim promocijskim slogom podjetij, javnih ustanov in nevladnih organizacij, kar nakazuje na potrebo po natančnejši kategorizaciji korporativnih računov v korpusu, ki jo načrtujemo za prihodnje raziskave.

Pričujoči prispevek je dragocen tudi zato, ker demonstrira potencial korpusnih pristopov v komunikologiji, medijskih študijah in drugih sorodnih družboslovnih disciplinah, ki temeljijo na jezikovnih podatkih, kar v slovenskem okolju še ni ustaljena praksa. Poleg teoretične relevantnosti imajo rezultati predstavljene analize tudi praktično vrednost za komunikološko stroko, saj izpostavljajo pomen ustrezno usposobljenih strokovnjakov za odnose z javnostmi, ki obvladajo dialoški, simetričen model družbenih omrežij, razumejo svojo posredniško vlogo med deležniki in podjetjem, ki ga zastopajo, ter proaktivno iščejo priložnosti za navezovanje pristnih stikov z deležniki in spodbujajo dialog z njimi.

1.01 UDC: 003.295: 342.537.6(497.4)"2014/2018"

Darja Fišer,[*] Nikola Ljubešić,[**] Tomaž Erjavec[***]

Parlameter – a Corpus of Contemporary Slovene Parliamentary Proceedings

IZVLEČEK

PARLAMETER – KORPUS RAZPRAV SLOVENSKEGA DRŽAVNEGA ZBORA

V prispevku predstavimo korpus sodobnih parlamentarnih razprav Parlameter, ki vsebuje razprave 7. mandata slovenskega Državnega zbora (2014–2018). Korpus Parlameter vsebuje bogate metapodatke o govorcih (spol, starost, izobrazba, strankarska pripadnost) in je jezikoslovno označen (lematizacija, tegiranje), kar omogoča številne raziskave s področja digitalne humanistike in družboslovja. V prispevku prikažemo potencial korpusnoanalitičnih tehnik za raziskovanje političnih razprav. Korpusna arhitektura je zasnovana tako, da omogoča širitev korpusa na druga časovna obdobja, prav tako pa tudi vključevanje gradiv drugih parlamentov, začenši s hrvaškim in bosanskim.

Ključne besede: parlamentarne razprave, izdelava korpusa, jezikovne tehnologije, korpusna analiza

ABSTRACT

The paper presents the Parlameter corpus of contemporary Slovene parliamentary proceedings, which covers the VIIth mandate of the Slovene Parliament (2014–2018). The Parlameter corpus offers rich speaker metadata (gender, age, education, party affiliation)

* **Department of Translation, Faculty of Arts, University of Ljubljana, Aškerčeva cesta 2, SI-1000 Ljubljana, Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, darja.fiser@ff.uni-lj.si**

** **Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, nikola.ljubesic@ijs.si**

*** **Department of Knowledge Technologies, Jožef Stefan Institute, Jamova Cesta 39, SI-1000 Ljubljana, tomaz.erjavec@ijs.si**

and is linguistically annotated (lemmatization, tagging), which boost research in several digital humanities and social sciences disciplines. We demonstrate the potential of the corpus analysis techniques for investigating political debates. The corpus architecture allows for regular extensions of the corpus with additional Slovene data, as well as data from other parliaments, starting with Croatian and Bosnian.

Keywords: parliamentary proceedings, corpus construction, language technology, corpus analysis

Introduction

Parliamentary discourse is motivated by a wide range of communicative goals, from position-claiming, persuasion and negotiation to agenda-setting and opinion-building along ideological or party lines. It is characterized by role-based commitments and confrontation and the awareness of a multi-layered audience (Ilie 2017). The unique content, structure and language of records of parliamentary debates are all factors that make them an important object of study in a wide range disciplines in digital humanities and social sciences, such as political science (van Dijk 2010), sociology (Cheng 2015), history (Pančur and Šorn 2016), discourse analysis (Hirst et al. 2014), sociolinguistics (Rheault et al. 2016), and multilinguality (Bayley 2014).

Despite the fact that parliamentary discourse has become an increasingly important research topic in various fields of digital humanities and social sciences in the past 50 years (Chester and Bowring 1962; Franklin and Norton 1993), it has only recently started to acquire a truly interdisciplinary scope (Bayley 2014). Recent developments enable cross-fertilization of linguistic studies with other disciplines and in-depth exploration of institutional uses of language, interpersonal behaviour patterns, interplay between language-shaped facts, and reality-prompted language ritualization and change (Ihalainen et al. 2016).

With an increasingly decisive role of parliaments and their rapidly changing relations with the public, mass media, executive branch and international organizations, further empirical research and development of integrative analytical tools are necessary in order to achieve a better understanding of parliamentary discourse as well as its wider societal impact, in particular with studies that represent diverse parts of society (women, minorities, marginalized groups) and cross-cultural studies (Hughes et al. 2013).

Parliamentary Corpora

The most distinguishing characteristic of records of parliamentary debates is that they are essentially transcriptions of spoken language produced in controlled and regulated circumstances. For this reason, they are rich in invaluable (sociodemographic)

meta-data. They are also easily available under various Freedom of Information Acts set in place to enable informed participation by the public and to improve effective functioning of democratic systems, making the datasets even more valuable for researchers with heterogeneous backgrounds.

This has motivated a number of national as well as international initiatives (for an overview, see Fišer and Lenardič 2018) to compile, process and analyse parliamentary corpora. They are available for most countries within the CLARIN ERIC research infrastructure for language resources and technology, with the UK's Hansard Corpus being the largest (1.6 billion tokens) and spanning the longest time period (1803–2005) while corpora from other countries are significantly smaller (most comprise between 10 and 100 million tokens) and cover significantly shorter periods (mostly from the 1970s onwards).

The Slovene parliamentary corpus SlovParl 2.0 (Pančur 2016) contains minutes of the Assembly of the Republic of Slovenia for the legislative period 1990–1992 when Slovenia became an independent country. The corpus comprises over 200 sessions, almost 60,000 speeches and 11 million words. It contains extensive meta-data about the speakers, a typology of sessions and structural and editorial annotations and is uniformly encoded to the Text Encoding Initiative (TEI) Guidelines, a de-facto standard for encoding and annotating textual data in Digital Humanities. It is available under the CC-BY licence in the CLARIN.SI repository of language resources and via the CLARIN.SI concordancers (Pančur et al. 2017). SlovParl is thus an exemplary corpus but contains material from a quite limited, and not very recent time period. This makes the corpus of limited use for the rich body of research on recent parliamentary activities.

Contemporary Slovenian parliamentary debates are monitored by the analytical tool Parlameter[11] which makes use of linguistic as well as non-linguistic data, such as MPs' attendance and voting results. While this is a very useful tool for journalists and citizen scientists and gives valuable insight into contemporary parliamentary data, its functionality is confined to that of the tool and as such cannot be freely manipulated by scholars according to their specific research needs.

The goal of the research presented in this paper was to convert the Parlameter database into a freely and openly available linguistically annotated corpus enriched with session and speaker metadata, and to showcase the analyses that can be performed on such corpora via open-source tools for corpus analysis. Section 3 gives the basic information on the corpus structure and size, Section 4 presents the analysis of the corpus according to the text and speaker metadata by utilizing some of the best-known corpus analysis techniques, and Section 5 gives some conclusions and directions for further research.

While the focus of the paper is the parliamentary language material which we process with natural language processing and analyse with standard methods from corpus linguistics, the aim of the analysis is to inform media and political studies by transferring the presented methodology into these areas.

1 *Parlameter*, https://parlameter.si.

Corpus Compilation

The data dump from the Parlameter tool consisted of the minutes of the National Assembly of the Republic of Slovenia from its VIIth mandate spanning sessions that started from 2014-08-01 to 2018-05-24 (the complete mandated lasted till 2018-06-22). It was received from the Parlameter API (application programming interface) as a series of JSON files, which were first reorganised into a file containing speaker metadata and a file with the transcriptions of the minutes with speaker identifiers. The speaker metadata contains information about the speaker name and surname, and (for some speakers) their sex, date of birth, education, and party affiliation. The complete speaker metadata is available for the members of the parliament and of the government, but not for, e.g., visiting field experts, representatives of governmental agencies, non-governmental organizations or civil initiatives. This is why the analyses in Section 4 are performed based on the instances for which the metadata is available in the corpus.

The transcriptions contain the ID of the session, name of the session (e.g. "*4. izredna seja*" - *4th extraordinary session*), the date when the session started, and its speeches, each one with the ID of the speaker and a number of segments, roughly corresponding to paragraphs. As discussed below, the transcriptions also contain comments by the transcribers.

Normalisation of Speaker Data

The speaker data was normalised by removing extraneous spaces and removing honorifics (sometimes the name was preceded by, e.g., "*Gospod*" – *Mr.*). Furthermore, in Slovene it is relatively easy to infer the sex from the given name, so we also added sex information to the speakers missing it.

Normalisation of Transcriptions

The JSON dump also contained empty speeches, as well as a significant amount of duplicated speeches. These were removed, as well as extraneous spaces in the text of the transcriptions.

Second, apart from the speeches, the minutes also contained 65,965 comments on verbal and non-verbal behaviour of the speaker or the members of parliament, and there are two types of such remarks. The first are written between slashes and are mostly comments on audible incidents, e.g., */nerazumljivo/ (incomprehensible), /oglašanje iz dvorane/ (comments from the hall), /znak za konec razprave/ (sign for the end of the discussion)*. The second type of comments are written between brackets and mainly denote voting results, e.g., *(nihče), /nobody/, (10 članov) /10 members/, (proti 44) /44 against/*. Both types of comments have been removed from the transcriptions

for the current version of the corpus, as they are not part of the transcription proper and would significantly complicate further processing. Furthermore, the content of the comments is not uniform, with the same information written in various ways (e.g. */smeh/ – laughter, /smeh iz dvorane/ – laughter from the hall, /smeh v dvorani/ – laughter in the hall*), meaning that the values would have to be unified before being converted to appropriate corpus elements.

Linguistic Annotation

In the second stage, the text of the transcriptions was automatically annotated with linguistic information. In particular, the text was tokenised, i.e. split into words, punctuation marks and spaces, and segmented into sentences, which was performed by the ReLDI tokeniser (Ljubešić et al. 2016). Second, the words were part-of-speech tagged and lemmatised, i.e. each word was assigned its context-dependent morphosyntactic description and non-marked form, e.g., the words in "*V naši sredini*" – *In our midst* are assigned the MSDs "*Sl Ps1fslp Ncfsl*" meaning preposition in the locative case; the possessive pronoun in the first person feminine singular locative with a plural owner number; and the feminine common noun in the singular locative, while the lemmas are "*v naš sredina*". The tagging and lemmatisation was performed with the ReLDI tagger (Ljubešić and Erjavec 2016) using its model for Slovene. Finally, the transcriptions were also tagged for named entities, i.e., names identified in the corpus were marked and categorised into five classes, those for persons, locations, organisations, for adjectives derived from a person's name (e.g. "*Cerarjev*" – *Cerar's*), and a miscellaneous category. The named entity annotation was performed with Janes-NER (Fišer et al. 2018).

Corpus Encoding

The corpus is encoded in XML, according to the Text Encoding Initiative Guidelines (TEI Consortium 2017). The complete corpus is stored as one TEI document, which contains its TEI header with the metadata for the corpus, and its text body, containing the transcriptions, one division for each starting date of the sessions; each division is stored as a separate file, giving one root file for the corpus and 525 files for the divisions.

The TEI header contains extensive metadata for the corpus as a whole, e.g., its authors and funders, the source description, the list and numbers of elements used in the corpus, as well as the list of speakers and their metadata. Most metadata is given both in Slovene and English.

As illustrated in Figure 1, the TEI text body date divisions contain a division for each session, and then the utterances for each speaker, each one containing one or more segments, which then contain the annotated transcription.

Figure 1: The TEI encoding of the corpus.

```
<div xmlns="http://www.tei-c.org/ns/1.0" type="date">
 <docDate when="2014-08-26">26.08.2014–</docDate>
 <head>Mandat VII, 26.08.2014–</head>
 <div type="session">
  <head>2. redna seja</head>
  <docDate when="2014-08-26">26.08.2014–</docDate>
  <u xml:id="u529092" who="#spk11">
   <seg xml:id="u529092.seg1">
    <s xml:id="u529092.seg1.1">
     <w lemma="lepo" ana="mte:Rgp">Lepo</w><c> </c>
     <w lemma="pozdravljen" ana="mte:Appmpn">pozdravljeni</w>
     <pc ana="mte:Z">.</pc><c> </c>
    </s>
    <s xml:id="u529092.seg1.2">
     <w lemma="pričenjati" ana="mte:Vmpr1p">Pričenjamo</w><c> </c>
     <w lemma="2." ana="mte:Mdo">2.</w><c> </c>
     <w lemma="seja" ana="mte:Ncfsa">sejo</w><c> </c>
     <w lemma="kolegij" ana="mte:Ncmsg">Kolegija</w><c> </c>
     <w lemma="predsednik" ana="mte:Ncmsg">predsednika</w><c> </c>
     <name type="org">
      <w lemma="državen" ana="mte:Agpmsg">Državnega</w><c> </c>
      <w lemma="zbor" ana="mte:Ncmsg">zbora</w>
     </name>
     <pc ana="mte:Z">.</pc>
    </s>
```

Corpus Size

Some basic statistics regarding the corpus are given in Table 1. In total, the Parlameter corpus contains 371 sessions (as distinguished by their title) which spanned over 525 days, i.e., 1.4 days per session on average. If we count distinct sessions that started on a given day, the corpus contains 1,338 such sessions. The VIIth mandate of the parliament heard 1,981 speakers who gave 133,287 speeches which contain almost 35 million words, i.e., 67 speeches per speaker and 260 words per speech on average. Due to a number of factors, such as different roles of the speakers in the parliament, the distribution is, of course, far from uniform, e.g., there is one speaker that gave 14,616 speeches, while 711 speakers gave only one speech.

Table 1: Basic statistic of the Parlameter corpus.

Tokens	40,987,516
Words	34,882,499
Sentences	1,833,147
Utterances	133,287
Speakers	1,981
Sessions on date	1,338
Dates	525
Sessions	371

Availability of the Corpus

The Parlameter corpus is available through CLARIN.SI. CLARIN is the European research infrastructure for language resources and technologies, which makes digital language resources available to scholars, researchers, students and citizen-scientists from all disciplines, especially in the humanities and social sciences, through single sign-on access. CLARIN offers long-term solutions and technology services for deploying, connecting, analysing and sustaining digital language data and tools. CLARIN is organised as a network of national centres, with CLARIN.SI covering Slovenia. CLARIN.SI[2] offers, inter alia, two concordancers for on-line corpus exploration, and a repository of language resources and tools, intended for their long-term archiving together with support for different types of licences and an unambiguous way for others to cite these resources, using Handle persistent identifiers. The landing page of each resource also gives a cross-reference to the concordancers for the particular corpus, and vice-versa. The repository also exposes its metadata, which is being harvested by a number of other services.

The Parlameter corpus is available through both CLARIN.SI concordancers, as well as for download from its repository, both as a TEI document and in the simpler vertical file format, under the liberal Creative Commons – Attribution-ShareAlike (CC BY-SA 4.0) licence (Dobranić et al. 2019). In this way we hope to raise interest among other researchers to explore the corpus and make use of it in their research.

2 *CLARIN Slovenia*, http://www.clarin.si/info/about/.

Corpus Analysis

By using the CLARIN.SI NoSketch Engine concordancer,[3] we demonstrate the potential of the basic corpus analysis techniques (Gorjanc and Fišer 2013) for politology, history and other related humanities and social sciences disciplines that base their research on large volumes of language data. *Concordances* are lists of all examples of the search word or phrase from a corpus which are shown in the context they were used in and are equipped with the available metadata. *Wordlists* are comprehensive summarizations of the language inventory in the corpus, organized by frequency or alphabetically. *Collocations* are partly or fully fixed multi-word expressions which have become established through usage. *Keywords* are words which appear in the focus corpus more frequently than they would in the general language. Combined with the available text and speaker metadata, such as date, speaker gender or political affiliation, they provide a powerful analytical tool for discovering the commonalities and specificities of the linguistic footprint and trends by different types of speakers in the parliament as will be shown in the rest of this section.

Production Volume and Vocabulary Size

As already presented in Table 1, the corpus contains nearly 41 million tokens or 35 million words. noSketch Engine also offers the lexicon size of the corpus, as given in Table 2, which shows that the corpus contains approximately 263,000 different word forms (so, inflected words, e.g., *Slovenije*) and over 104,000 different lemmas (so, base forms of words, e.g., *Slovenija*), and 1,080 different morphosyntactic tags (e.g.,*Verb main present second plural*). However, it should be noted that both lemmas and the tags are automatically assigned, so they also contain some annotation errors: the accuracy of morphosyntactic tags is around 94%, the accuracy of lemmas is above 99%.

Table 2: Lexicon sizes of the Parlameter corpus.

Unique words	263,007
Unique lemmas	104,247
Unique tags	1,080

While the corpus contains parliamentary debates from the period 2014-2018, 62% of the material was recorded in 2015 and 2016. Given the parliamentary term, which lasted from 1 August 2014 to 14 April 2018, it is interesting to observe an 8% smaller production in 2017 compared to the year before since the last year of the term would be expectedly the busiest in order to wrap up the workplan and set the ground for a new election cycle.

3 *NoSketch Engine @ CLARIN.SI*, https://www.clarin.si/noske/.

Table 3: Distribution of text quantity by year in Parlameter.

Year	No. of tokens	% of tokens	Rel. freq.
2014	3,759,110	9%	91,714
2015	12,441,754	30%	303,550
2016	13,270,257	32%	323,763
2017	9,944,401	24%	242,620
2018	1,571,994	4%	38,353
Total	40,987,516	100%	1,000,000

Morphosyntactic Specificities of the Language in ParlaMeter

We performed a basic analysis of the morphosyntactic annotations of the corpus in form of the most significant differences in their frequencies between the Gigafida reference corpus of Slovene[4] and the Parlameter corpus, which are given in Table 4.[5]

Table 4: Most salient differences in morphosyntactic descriptions between Gigafida 2.0 and Parlameter.

Gigafida	Parlameter
Residual web	Pronoun personal first singular nominative
Numeral roman cardinal	Verb main present second plural
Adjective possessive positive masculine singular instrumental	Pronoun personal second masculine plural nominative
Auxiliary infinitive	Pronoun possessive first feminine singular genitive singular
Adjective possessive positive masculine plural genitive	Verb main present first plural -Negative
Adjective possessive positive masculine singular locative	Verb main present second plural -Negative
Adjective possessive positive neuter singular locative	Pronoun demonstrative neuter plural accusative
Pronoun possessive third masculine singular accusative dual	Pronoun personal first singular accusative
Adjective possessive positive masculine singular nominative -Definiteness	Verb main present first singular

4 For this comparison we used the deduplicated version of Gigafida 2.0. At the time of writing, this corpus was newly made and does not yet have a reference publication. It is, however, freely available for searching and analysis at https://www.clarin.si/noske/.

5 The morphosyntactic tags are given here in their expanded form to aid understanding. The reference to these morphosyntactic descriptions is given in http://nl.ijs.si/ME/V6/msd/html/msd-sl.html.

Gigafida	Parlameter
Pronoun possessive third feminine plural locative singular masculine	Verb main present first singular
Adjective possessive positive masculine plural nominative	Pronoun demonstrative masculine singular dative
Noun proper feminine plural dative	Pronoun indefinite feminine singular genitive
Numeral letter ordinal neuter plural genitive	Pronoun indefinite masculine singular accusative
Pronoun personal first dual accusative	Verb auxiliary present second plural -Negative
Pronoun personal first dual dative	Verb auxiliary future first singular -Negative
Noun proper neuter singular instrumental	Pronoun personal first masculine plural nominative
Adjective possessive positive feminine singular locative	Verb auxiliary present second plural +Negative
Pronoun personal second singular accusative bound	Verb main present first plural
Pronoun personal third masculine dual dative +Clitic	Pronoun indefinite feminine singular accusative
Adjective possessive positive masculine plural locative	Pronoun demonstrative feminine plural accusative

The results show that the parliamentary speeches, as expected, contain more present tense verb forms, especially in the first and second person singular or plural (e.g., *imamo – we have, pozdravljam – I greet, zaupate- you trust*), as well as personal and demonstrative pronouns, the former most prominently as the first person singular personal pronoun *(jaz – I)*.

On the other hand, the parliamentary proceedings do not contain URLs or Roman numerals. More interestingly, they also contain significantly fewer possessive adjectives (e.g. *torkovim – Tuesday's*) and pronouns (*njun – theirs*$_{[dual]}$), proper names, numerals, personal pronouns in the dual number (*naju – us two*), or in second person singular accusative (*nate – to you*) than general Slovene.

Language and Gender in Parlameter

Gender is recorded for all but one speaker in the corpus.[6] In total, 1,965 speakers are represented, 62% of which are male and 38% female. Interestingly, the contribution from the speakers is not proportionate to the distribution according to their gender,

6 This missing information is due to errors in input metadata records, which will be improved in the next version of the corpus.

with the male speakers contributing 71% of the tokens in the corpus and the female speakers 29%. On the speech level the difference is even more pronounced as the male speakers delivered 73% of the speeches while female speakers only 27%, indicating that, on average, the speeches given by female speakers were somewhat longer than those by male speakers.

Table 5: Distribution of speakers and text production by gender in Parlameter.

Gender	No. of speakers	% of speakers	No. of tokens	% of tokens
Female	747	38%	29,147,871	71%
Male	1217	62%	11,838,913	29%
Unknown	1	0%	732	0%
Total	1965	100%	40,987,516	100%

Table 6, which lists top-ranking 10 female and male speakers and their production in terms of tokens, shows that the most prolific male speakers produced nearly twice as much material as their female counterparts. Overall, all top 10 speakers except one (Miha Kordiš, male, the Levica party) have a leading role in one or more parliamentary or governmental bodies, including 2 ministers, both of which are female, 2 opposition deputy group chairs, who are both male, and the Chair of the National Assembly who is also male. Based on their roles in the parliament or the government, top-ranking speakers represent issues on culture, corruption, judiciary, finances, agriculture, foreign policy, education and infrastructure. In terms of political orientation, the largest opposition party SDS is best represented with 5 top-ranking male and 3 female speakers, including chair and vice-chair of their deputy group. Among the top-ranking female speakers, the entire political spectrum is represented while male speakers from the SD and DeSUS parties do not make the list, and the SMC party is only represented by the Chair of the National Assembly whose role is most likely predominantly procedural, not to promote the party agenda.

Table 6: Top-ranking 10 female and male speakers and their text production in Parlameter.

Female	Party affiliation // Role	Tok. %	Male	Party affiliation // Role	Tok. %
Anja B. Žibert	SDS // Chair of the Culture Committee	698,883 6%	Jožef Horvat	NSI // Chair of the Foreign Policy Committee; Chair of the Deputy Group NSI	1,141,778 4%
Jelka Godec	SDS // Chair of the Inquiry Commission on the Misuse Practices in Healthcare	530,029 4%	Jani Mödern-dorfer	ZAAB // Chair of the Inquiry Commission on bank money laundering; Vice-chair of the Election Committee	1,062,546 4%

Female	Party affiliation // Role	Tok. %	Male	Party affiliation // Role	Tok. %
Iva Dimic	NSI // Vice-chair of the Judiciary Committee	509,101 4%	Franc Trček	Levica // Vice-chair of the Infrastructure Committee; Vice-chair of the Inquiry Commission on bank money laundering	1,060,399 4%
Alenka Bratušek	ZAAB // Vice-chair of the Public Finances Committee; Vice-chair of the Deupty Group ZAAB	483,171 4%	Milan Brglez	SMC // Chair of the National Assembly; Chair of the Constitution Committee	948,334 3%
Violeta Tomić	Levica // Vice-chair of the Agriculture Committee	446,460 4%	Vinko Gorenak	SDS // Vice-chair of the Deputy Group SDS	788,678 3%
Eva Irgl	SDS // Chair of the petition committee	439,042 4%	Franc Breznik	SDS // Vice-chair of the Election Committee	763,437 3%
Urška Ban	SMC // Chair of the Finances and Monetary Policy Committee	382,425 3%	Jože Tanko	SDS // Chair of the Deputy Group SDS	752,130 3%
Mateja V. Erman	Minister of Finance	381,604 3%	Andrej Šircelj	SDS // Chair of the Public Finances Committee	721,135 2%
Bojana Muršič	SD // Vice-chair of the National Assembly, Vice-chair of the Education Committee	366,547 3%	Tomaž Lisec	SDS // Chair of the Agriculture Committee	707,666 2%
Julijana B. Mlakar	DeSUS // Minister of Culture; Vice-chair of the Foreign Policy Committee	308,355 3%	Miha Kordiš	Levica	676,717 2%

In order to compare the topics discussed by female and male speakers in the Slovene parliament, we analysed their 100 top-ranking key lemmas, where we used the corpus of all female speakers as the target corpus against the reference corpus of all male speakers in the Parlameter corpus, and vice versa, so the two lists display the distinguishing features of each of the groups. By observing their contexts via concordances, we manually classified them into one of the 13 topics represented by the ministries in the Slovenian government:

- *agriculture, forestry and food*
- *culture*
- *defence*
- *economy and technology*
- *education, science and sport*
- *environment and spatial planning*
- *finance*
- *health*
- *foreign affairs*
- *infrastructure*
- *interior*
- *justice*
- *labour, family and social affairs*
- *public administration*

In addition, we introduced 4 additional categories for words that could not be classified into any of the topics above:

- *interaction/procedural* for keywords which referred to other people attending the session (e.g., references to names of other speakers, *predsednik – chairman*) or expressed procedural matters during the session (e.g., *prisotni – present, dobrodošli – welcome*)
- *style* for keywords which were either distinctly colloquial or distinctly formal and were frequently used only by a single or very few speakers in order to achieve a special effect (e.g., *penez*, a very informal expression for money, *šiht*, a very informal expression for job)
- *ideology* for keywords which were used to ideologically label an individual speaker or a group of speakers (e.g., *levičarski – leftist, kapitalizem – capitalism*)
- *multiple* for keywords which were used in several topics (e.g., *zgodnji – early, fantastičen – fantastic*).

As can be seen from Table 7, the most frequent topics among the female speakers are *health* (35) and *labour, family and social affairs* (33), which are followed by *public administration* (13) and *education, science and sport* (8). Most of the 100 top-ranking keywords uttered by male speakers, on the other hand, could not be classified into a single topic because they were used either to achieve a *stylistic effect* (24), were general words that were used in *multiple topics*, such as descriptive adjectives or legal terms (22), or *ideological expressions* (6), all of which indicate a more discursive, debating style of the male speakers, but could also stem from the fact that the leading roles in that term were predominantly held by male members of parliament.[7] Despite being much more infrequent than in the female part of the corpus overall, the most

7 This problem could be avoided by removing outliers regarding production in the dataset before performing the analyses. But our goal here was to present the complete corpus and demonstrate the basic corpus analysis techniques.

frequently represented specific topics by male speakers are *infrastructure* (9), *interior* (6), *agriculture, forestry and food* (5), and *defence* (5), suggesting a significant difference in the roles and interests of male and female speakers in the Slovene parliament.

Table 7: Topics of 100 top-ranking keywords of female and male speakers in Parlameter.

Topics – female	Freq.	Topics – male	Freq.
health	35	style	24
labour, family & social affairs	33	multiple	22
public administration	13	infrastructure	9
education, science & sport	8	interior	6
interaction/procedural	3	ideology	6
multiple	3	interaction/procedural	5
environment & spatial planning	1	agriculture, forestry & food	5
agriculture, forestry & food	1	defense	5
culture	1	foreign affairs	4
finance	1	finance	4
economy & technology	1	justice	3
Total	100	Total	100

Illustrative examples of the 10 top-ranking female- and male-specific keywords with a manually assigned topic are listed in Tables 8 and 9.

Table 8: Most frequent keywords, topics and word type among female speakers in Parlameter. N stands for nouns, Adj for adjectives, and NP for proper nouns (names).

Lemma – English translation	Topic	PoS	Freq.	Freq_ref	Score
rejništvo – fostercare	labour, family & social affairs	N	264	59	7.7
mark – mark	health	PN	155	29	7.1
enostarševski – single-parent	labour, family & social affairs	Adj	167	38	6.6
roditeljski – parent	labour, family & social affairs	Adj	169	39	6.5
medical – medical	health	PN	128	26	6.2
plazma – plasma	health	N	82	9	6.1
pacientov – patient's	health	Adj	282	97	5.7
zaznamba – notice	public administration	N	155	43	5.7
žilen – stent	health	Adj	518	213	5.4
duševen – mental	health	Adj	393	156	5.4
nasilnež – violent person	labour, family & social affairs	N	98	21	5.4

Table 9: Most frequent keywords, topics and word type among male speakers in Parlameter.

lemma – English translation	category	PoS	Freq_ref	Score
penez – inf. money	finance	N	0	13.2
navsezadnje – nevertheless	multiple	Adv	90	8.4
kubik – cubic	agriculture, forestry & food	N	10	7.8
islam – Islam	interior	N	6	6.4
levičarski – leftist	ideology	Adj	2	6.2
navzoč – present	interaction/procedural	Adj	211	6.0
avtošola – driving school	infrastructure	N	1	5.8
socialist – socialist	ideology	N	25	5.5
svojevrsten – peculiar	multiple	Adj	16	5.4
e-klopa – e-bench	interaction/procedural	N	1	5.3
prečenje – crossing	style	N	3	5.2

That the nature and style of male speeches is quite different from the female ones can also be seen from the analysis of the morphosyntactic types of 100 highest-ranking keywords for male and female speakers. While nouns are the most frequent category and are used equally frequently by both male and female speakers (44%), many more adjectives were found among the female top-ranking keywords (33% vs. 16%), while the male keywords had more adverbs (11% vs. 4%) and verbs (9% vs. 2%), which again could be related to the roles of the speakers in the parliament. However, given the results of our preliminary work on this dataset (Ljubešić et al. 2018), during which we removed the speakers that produced most of the linguistic material from the analysis, we see similar trends both in the gender-dependent keyword and morphosyntactic analysis, and are therefore rather in favour of accepting the observed differences as impact of gender and not role.

Language and Party Affiliation in Parlameter

Affiliation is recorded for only 113 speakers out of the 1982, however, these are responsible for 79% of the tokens in the corpus. Affiliation is considered as either deputy group membership or a role in the government, where it must be noted that in this version of the corpus the metadata reflect the situation at the beginning of the term and does not keep track of party membership transfers or resignations of ministers or members of parliament. Also, when elected members of parliament were later appointed as ministers, the metadata record only their party affiliation and records as ministers only those who were appointed without being first elected to the parliament. To facilitate more fine-grained and accurate use of the corpus in political science or contemporary history, we plan to refine the metadata for the next release of the corpus, adding also the MP's membership in the working bodies of the National Assembly,

etc. Also, the metadata in the current version of the corpus do not flag the independent members of parliament who do not belong to any of the parliamentary parties and operate in the Independents deputy group, which is why they are not included in our analysis.

As Table 10 shows, the most prolific deputy group is the largest opposition party Slovenian Democratic Party (SDS), whose 20 members contributed nearly 10 million tokens or 30% of the corpus. SDS is followed by the main governing party, Party of Modern Centre (SMC), whose 42 members contributed 7 million tokens or 22% of the corpus. It is interesting to note that in terms of the volume contributed to the corpus on one side and the number of speakers on the other, that this party was the least productive among the main parties, with a ratio of the percentage of tokens to the percentage of speakers (i.e., the relative token to speaker ratio) of 0.54, which means that this party generated a little bit more than a half of the material that would have been expected given their number of speakers and the overall activity of all the speakers. The Left (Levica) and New Slovenia (NSi) rank third and fourth, despite the fact that they had only 6 members each in the parliament, making them the most productive parties with a relative token to speaker ratio of 1.83 and 1.66. The Democratic Party of Pensioners of Slovenia had as many as 12 elected MPs but contributed 1 million tokens less than the two previous parties, which makes them the second least productive party with a relative token to speaker ratio of 0.67.

Table 10: Distribution of speakers and text production by party affiliation in ParlaMeter with speakers with unknown affiliation removed.[8]

Affiliation	No. of speakers	% of speakers	No. of tokens	% of tokens
Slovenian Democratic Party Deputy Group (SDS)	20	20%	9.516.651	30%
Party of Modern Centre Deputy Group (SMC)	42	41%	7.162.719	22%
The Left Deputy Group (Levica)	6	6%	3.438.194	11%
New Slovenia – Christian Democrats Deputy Group (NSI)	6	6%	3.370.131	10%
Social Democrats Deputy Group (SD)	9	9%	2.533.019	8%
Democratic Party of Pensioners of Slovenia Deputy Group (DeSUS)	12	12%	2.435.884	8%
Party of Alenka Bratušek Deputy Group (SAB)	4	4%	1.876.294	6%
Italian and Hugarian National Minorities Deputy Group (IMNS)	2	2%	117.709	0%
Government	1	1%	1.765.374	5%
Total	102	100%	32.215.975	100%

8 The number of speakers per party is calculated from the ParlaMeter dump and deviates slightly from the official member number due to different handling of speakers with multiple roles.

Next, we performed a manual analysis of the 100 top-ranking keywords of each political party against the rest of the corpus. These analyses display the distinct properties of one party that are not shared by other parties. Using the concordances, we classified the keywords into the same categories as in Section 4.1, the results of which are summarized in Tables 11 and 12.

Table 11: Topics of 100 top-ranking keywords of party members in Parlameter.

Topics	SMC	DeSUS	SD	SDS	NSi	Levica	SAB
agriculture, forestry & food	0	0	**34**	0	**27**	0	0
culture	0	3	0	0	0	1	0
defense	0	0	**21**	5	0	0	1
economy & technology	0	0	5	1	**11**	**13**	1
education, science & sport	0	0	0	0	0	0	4
environment & spatial planning	0	0	3	0	6	1	0
finance	0	2	2	0	6	1	1
foreign affairs	0	5	0	2	4	3	0
health	0	3	0	8	1	0	5
ideology	0	0	0	**15**	3	9	0
infrastructure	1	0	2	0	7	1	1
interaction/procedural	**99**	**61**	14	**17**	10	4	14
interior	0	0	0	3	0	3	5
justice	0	1	1	8	0	0	0
labour, family & social affairs	0	**13**	3	1	4	13	3
multiple	0	2	6	13	8	17	29
public administration	0	2	0	5	2	1	7
style	0	8	9	**22**	11	**33**	**29**
Total	100	100	100	100	100	100	100

Unsurprisingly, due to the role of the main governing party SMC, practically all their top-ranking keywords are interactional elements with the other speakers or have a procedural nature (e.g., *navzoč – present, glasovanje – voting, amandma – amendment*). That DeSUS is a single-issue party can be seen from their keywords, which, apart from a surprisingly high proportion of interactive keywords, belong almost exclusively to the semantic field of retirement and pension (e.g., *regres – holiday pay, valorizirati – to revalue, gmoten – material*). Interestingly, even the topics of foreign affairs and culture are nearly completely absent from their keyword list, despite the fact that these ministers came from their party, suggesting that these topics are more or less evenly shared with other parties. SD, the third coalition party, clearly display their priority areas of agriculture, forestry and food (e.g., *teran – Teran wine, fermentiran – fermented, kmetovati – to farm*) and defence (e.g., *vojakinja – female soldier, neeksplodiran – unexploded, strelivo – ammunition*), which can be traced back to their ministers.

The largest opposition party SDS stands out from the rest by the amount of ideological keywords identified among the top-ranking keywords (e.g. *tranzicijski – transitional, totalitarizem – totalitarism, lustracija – lustration*). NSi and Levica, the opposition parties with the same number of MPs but from the opposite ends of the political spectrum, both address the widest variety of issues (their keywords were classified into 13 out of 18 topics). The topics with nearly equal number of completely opposite keywords are economy and technology (e.g. *soupravljanje – co-management* for Levica vs. *espejevec – private entrepreneur* for NSi). While NSi mostly talks about the local issues related to their constituencies (e.g. *samooskrba – self-sufficiency, posekan – cut down, obdelovati – farm*), Levica stands out by signature stylistic devices which range from very informal (e.g. *šlamastika – pickle, gazda – informal for master, nabijati – to bang on*) to highly elevated registers (e.g. *nemara – perhaps, onkraj – beyond, ducat – dozen*) and displays the largest proportion of ideological vocabulary next to SDS (e.g. *tovarišica – camerade, revizionizem – revisionism, imperializem – imperialism*). SAB seems to stand out by a predominantly (local) administrative/procedural/governance vocabulary (e.g. *proporcionalen – proportional, odpoklic – recall, dvokrožen – double-ballot*) as well as a discursive, informal style of distinctly negative sentiment, which is characteristic of one of their members Vinko Möderndorfer (e.g. *rešpektiram – honour, kozlarija – nonsense, zmazek – disaster*).

Table 12: 100 top-ranking keywords per political party, taking into account lowercased lemmas, computed against the rest of the Parlameter corpus and sorted according to their keyness score.

SMC	navzoč, e-klopa, udis, roberto, prekinjen, podprogram, prehajati, lipicer, kustec, katerim, grebenšek, h, battelli, epi, stanujoč, obveščati, krajnc, zaključevati, predajati, pričenjati, sodin, porotnica, simona, franc, glasovati, obrazložitev, moderen, kolegij, tanko, postopkovno, potisek, končevati, nuklearen, brezpredmeten, ep, jernej, dneven, počkaj, glasovnica, mandatno-volilen, vojko, jožef, trček, bojan, neusklajen, tilen, prelog, ustavnorevizijski, odločanje, arko, nadomeščati, he, branislav, matej, jože, glasovanje, prvopodpisan, e-klop, glas, dopolnjen, porotnik, terminski, vložen, simono, franca, pogačnik, erman, ugotavljati, klanjšček, smc, stebernak, nepovezan, jana, žibert, bien, matjaž, šircelj, fajt, postopkoven, lilijana, skrajšan, monetaren, prekinjati, poslovniški, matičen, bah, mag., marinka, šergan, lenča, vraničar, izvolitev, karlovšek, razpravljavec, predstavnica, razširitev, anita, amandma, nadomeščanje, zame
DeSUS	meglič, črnak, pripadajoč, desus, pogačar, dasiravno, vukov, valenca, požun, inferioren, upajoč, möderndorfer, pregrešiti, divjak, valorizacija, korva, rezime, kkr, kuzmanič, marijan, upokojen, vuk, mehčati, pojbič, košnik, bližnjevzhoden, zaposlovalen, punkcija, žmavc, milojka, zaporedno, celarc, konzularen, xv., marija, kolar, bačič, erika, grošelj, rubelj, minski, lukić, rudarski, zadržanost, mirjam, godec, valorizirati, sng, tašner, kušar, brinovšek, invalid, zamrznitev, tedaj, dvoživkarstvo, nina, pirnat, dekleva, merše, federacija, nada, klanjšček, protiukrep, jelka, ogrizek, gmoten, kisikov, ivo, majcen, izvoliti, iva, dimic., modifikacija, ljubič, žan, upokojenec, prikrajšanje, prečitati, šimenko, jasna, izplačevanje, zipro, korpič, antonija, premožen, sapa, voljč, suzana, dimic, vesni, lukič, zdravko, irena, teja, sluga, regres, ruše, janja, razparava, trivialen
SD	izčistiti, genetsko, izčiščen, vezava, surov, demokrat, vojakinja, gorsko-hribovski, travinje, potočan, vadišče, razprodati, hip, služenje, hišniški, faktorski, pripadnica, stiskanje, zmogljivost, omd-, kočevski, anhovo, vrtojba, peterica, mineralen, maji, krušen, kmetica, ciolos, vklop, deti, socialdemokratski, formacijski, teran, selnica, kloniran, urszr, obramben, salonit, radeče, mlekarna, neperspektiven, marjana, popolnjevanje, omd, odzivanje, vrtnina, vselej, zorganizirati, vikariat, eutm, pokolp, govedo, rogaška, klirinški, razprodaja, surovina, ksenija, vinko, izčiščevati, konzumen, refundirati, pripadnik, neeksplodiran, social, uokviriti, žito, kfor, prebroditi, konvergenca, grajski, brecelj, hogan, administriranje, trader, kočevsko, h4, primož, korenjak, bržkone, kmetovati, obrtništvo, vojska, strelivo, poveljevanje, snežnik, plasiran, gorsko, refundacija, hribovski, proizvodnja, subvencijski, dacian, missing, kmetija, opazovati, voditeljstvo, kramar, fermentiran, viher
SDS	islam, fišer, mark, svinjarija, levičarski, odnosno, medical, kb, demokratski, odnosen, lenart, zemljarič, kučan, zalar, bordojski, kb1909, morišče, zločin, iznenada, velikanski, tomos, kangler, patria, multikulti, masleša, prvorazreden, škrlec, udba, stožice, tranzicijski, šef, praprotnik, moralno-etičen, ilegalno, zločinski, bomben, peticija, porsche, srebrenica, cener, umor, totalitaren, pokrasti, totalno, genocid, drugorazreden, tamle, erdogan, judikat, vega, ribičič, privilegiranec, komunističen, razorožitev, varnostnoobveščevalen, žilen, opornica, indičen, škandal, ornik, lustracija, poljanski, posavje, počenjati, furlan, pobiti, sevnica, ubog, janković, krkovič, npu, deček, opran, bojda, blamaža, lopov, toplak, kerševan, slikati, bmw, veselo, amen, totalen, komunizem, totalitarizem, obsoditi, preiskati, bedarija, udbovski, pomorjen, turnšek, vladavina, zlagati, šoping, vpiti, ukc, avion, klemenčič, koruptiven, neumnost

NSI	komunalno, socialno-tržen, marn, božičnica, zidanica, egalitaren, krščanski, espejevec, fantastičen, ekstrapolacija, planšarija, medparlamentaren, kamnik, demografija, kapica, bundestag, podonavski, bajuk, samoprispevek, vinogradnik, razlastiti, vipavski, prijateljstvo, kanalizacija, aksiom, pomurje, bogataš, ferenc, parcelacija, optimirati, oljčnik, komenda, polnost, vrtalec, ozp, pomurski, ikt, simulirati, dimniški, parlamentarec, podčrtovati, artikulirati, obžalovati, omizje, cerknica, polčas, ginijev, zbirno-reciklažen, brutalno, prekladanje, širokogruden, absorpcijski, šinko, dolenjsko, lestev, vodovod, rodnost, traktor, notranjska, opn, posekan, vinograd, zaraščati, odvajanje, loža, kristjan, davno, regresen, lovrenčič, firefox, parcela, akrapovič, obdelovati, obratovalnica, zpn, terezija, mihael, odlašati, peskovci, vamp, notranjski, ovs, copatek, veselica, upniški, penzija, hala, digitalen, goljuf, identifikacijski, mohar, postoriti, goveji, prirasti, splačati, samooskrba, prazniti, odstaven, todorić, pozor
Levica	penez, tuliti, vračljivost, ubesedovati, onkraj, bajta, neoliberalen, prečiti, nemara, ducat, socialist, delavski, imperialističen, zvrniti, desnica, navsezadnje, blazen, sociolog, šiht, soupravljanje, zategovanje, mandarin, kapitalizem, strokovec, šlamastika, blazno, kapitalističen, tovarišica, ubesedovanje, revizionizem, prekarnost, vzdržan, gazda, profit, sodržavljanka, izkoriščevalski, represija, protisocialen, nabijati, prekaren, metafora, soodločanje, periferen, agregaten, cinkarna, rezilen, mezda, amandmiranje, demokratizacija, ips, efektivno, natov, levica, belokranjec, bučka, zaposlovalec, izhajajoč, reven, požegnati, profiten, marof, ics, minimalec, podrejati, imperializem, kapitalist, silno, prekarizacija, odpustek, sodržavljan, noveliranje, versus, zvo, bolgarski, zastraševanje, informatičen, metaforično, režati, razreden, ciničen, striči, ropotati, korporacija, rasizem, redistributiven, pregrevanje, trade, rez, omv, prekeren, deregulacija, štacuna, grosist, znoreti, penzion, oligopolen, jahati, fevdalizacija, sočasno, prečenje
SAB	svojevrsten, večnost, mvk, pooblaščati, that's, diskvalifikacija, prekleto, bla, resnica, fakt, naglas, odpoklic, zavezništvo, minis, četrten, trapast, istrabenz, zasebništvo, zamah, dvokrožen, ramšakov, diskvalificirati, športnica, drk, štos, cetera, ups, nedostojno, redarski, strojan, nijz, proporcionalen, ma, evtanazija, zanič, bloudkov, etc, mv, vsakič, naturalizacija, zamera, nor, listnica, smešiti, dispečiranje, diskusija, strašansko, nefer, diskutirati, regres, sprevržen, r., zavrtanik, večen, hiv, nekorektno, ubežati, imperativen, presedan, prastrah, dinozaver, halo, ekstremističen, rimskokatoliški, mvk-, namenoma, zmazek, gedrih, somalijski, zamahniti, nonstop, kostanjevec, policaj, domišljati, prohibicija, znakoven, paradoks, barantati, et, hecen, močvirnik, avans, nametati, preprosto, prepričevati, podžupan, traparija, kričati, ekstra, non-stop, telovadba, stefanovič, el-zoheiry, ničkolikokrat, kozlarija, prvenstvo, boh, domišljija, rešpektiram

The Zeitgeist of ParlaMeter

Finally, we observe the zeitgeist of the Parlameter corpus by comparing it with its older and smaller cousin, the SlovParl corpus, which contains material from the period of Slovenia's independence (1990–1992). First, we created keyword lists with each of the two corpora acting as a focus and a reference corpus. We then manually classified 100 top-ranking keywords into the same categories as in Section 4.1, with the following additional categories:

– abbreviations (*etc., Mr.*), which were in use in the SlovParl but are no longer the convention in the ParlaMeter transcriptions of the parliamentary sessions
– IT vocabulary (*internet, web*), which at the time of SlovParl was not yet widespread.

If we disregard the differences in the mentions of the active politicians in the two periods, which are the most frequent category, most of the top-ranking keywords in both corpora belong to procedural and legal issues, which are clearly different in a newly established state and a state integrated in the EU (see Tables 13 and 14). Apart from that, many more topics are identified in the Parlameter corpus, such as economy and technology, foreign affairs and health, which again is not surprising as a well-established state will need to take care of a full spectrum of issues.

Table 13: Topics of the 100 top-ranking keywords in Parlameter and SlovParl.

Topic	ParlaMeter	SlovParl
abbreviation	0	3
defence	0	1
economy & technology	6	2
education	1	0
environment & spatial planning	2	0
finance	12	7
foreign affairs	4	0
health	4	0
multiple	0	1
informal vocabulary	2	0
infrastructure	1	0
interior	2	0
it vocabulary	2	0
justice	1	0
labour, family & social affairs	3	0
legal/procedural	14	21
politician/party	46	65
Total	100	100

Table 14: 100 top-ranking keywords in Parlameter contrasted against SlovParl and vice versa.

ParlaMeter	evro, eu, desus, smc, cerar, sdh, dutb, möderndorfer, trček, bratušek, sds, gorenak, spleten, mandatno-volilen, deležnik, koalicijski, kordiš, anja, matej, direktiva, postopkovno, kpk, okoljski, kohezijski, javnofinančen, tonin, bdp, veber, naročanje, korupcija, bah, jani, levica, nlb, unija, tanko, migrantski, povprečnina, vatovec, čakalen, pojbič, migrant, varuhinja, prikl, žnidar, šircelj, varuh, zujf, teš, violeta, tomić, mahnič, ddv, digitalen, han, istospolen, lisec, telekom, vrtovec, dars, žibert, novela, globa, zorčič, vajeništvo, godec, trošarina, čuš, okrožen, internet, prvopodpisan, schengenski, matić, trajnosten, gašperšič, jurša, podneben, dz, lipica, lah, podizvajalec, žan, uredba, blagajna, okej, verbič, ferluga, dobovšek, mramor, računski, vraničar, zakonik, ljudmila, nevladen, postopkoven, preiskovalen, direktorat, hanžek, muršič, irgl
SlovParl	delegat, oz., glavič, družbenopolitičen, gros, dinar, republiški, usklajevalen, din, skupščinski, starman, zakonjšek, alinea, vzdržati, potrč, vzdržan, kolešnik, izvršen, lukač, sklepčnost, pintar, npr., navzočnost, buser, arzenšek, feltrin, atelšek, liberalno-demokratski, smole, razpravljalec, školč, zvezen, schwarzbartl, delegatski, tomšič, zagožen, železarna, jakič, gošnik, skupščina, polajnar, tomažič, muren, štefančič, lastninjenje, deviza, zlobec, šter, demos, dretnik, kreditno-monetaren, sdp, čimprej, nabornik, devizen, marka, delegatka, sekretariat, bekeš, deželak, klavora, peterle, črnej, halb, kreft, šonc, lokar, gradišar, šeligo, juri, perko, sfrj, voljč, požarnik, semolič, volilec, kramarič, bučar, plebiscit, dvornik, tomše, grašič, tolar, starc, pregelj, podobnik, pozsonec, balažic, g., moge, medzborovski, jaša, razdevšek, rojec, šetinc, urbančič, lavtižar-bebler, vivod, anka, šešok

To illustrate differences in the zeitgeist of both corpora, we extracted the strongest collocations of the following 3 expressions, which are frequent in both corpora, taking into account the collocation candidates that appear at least 5 times immediately next (left or right) to the headword, and analysed the first 50 collocation candidates:

– adjective *južen – southern*,
– noun *kriza – crisis*, and
– verb *sprožiti – trigger*.

Table 15: Comparison of collocations of *južen*, *kriza* and *sprožiti* in SlovParl and ParlaMeter. Topics or morphosyntactic categories are indicated in square brackets, and new collocations in Parlameter are highlighted in bold.

	SlovParl	**ParlaMeter**
južen	178 (14.03 per million) - [GEOGRAPHY]: koreja, primorska, amerika - [CONCRETE]: meja, železnica - [METAPHORICAL]: trg, del, stran, republika	910 (22.20 per million) - [GEOGRAPHY]: **afrika**, koreja, **sredozemlje**, amerika, tirolska, sudan, **tirolec**, **koroška, italija, evropa, nemčija, slovenija** - [CONCRETE]: meja, **obvoznica, tok, sadje, odsek,** železnica, **ulica** - [METAPHORICAL]: **sosedstvo, soseda, sosed, soseščina,** del, trg, **projekt, stran, država, republika**
sprožiti	548 (43.19 per million) - [CONCRETE]: spor, postopek, proces, interpelacijo, arbitražo - [METAPHORICAL]: reakcijo, polemiko, akcijo, mehanizem, pobudo, vprašanje, diskusijo, zahtevo, spremembo, razpravo, zadevo	1,569 (38.28 per million) - [CONCRETE]: postopek, spor, **preiskavo, alarm,** process, **ovadbo, tožbo, stečaj, prijavo, revizijo** - [METAPHORICAL]: **plaz,** mehanizem, polemiko, reakcijo, **kepo,** pobudo, akcijo, **iniciativo, aktivnost, debato, kampanjo**
kriza	1,114 (87.79 per million) - [GEOGRAPHY]: jugoslovanska, zalivska kriza - [POLITICS]: vladna, gospodarska, parlamentarna, ekonomska, ustavna, politična kriza - [METAPHORICAL]: duševna, socialna, razvojna, družbena kriza - [MODIFIERS]: huda, moralna, globoka, katastrofalna, velika, težka kriza - [NOUNS]: reševanje, razrešitev, rešitev, razplet, razreševanje krize - [VERBS]: prebroditi, poglabljati, razrešiti, povzročiti, rešiti, začeti krizo	8,062 (196.69 per million) - [GEOGRAPHY]: ukrajinska, grška, svetovna, globalna kriza - [POLITICS]: **migrantska, begunska,** gospodarska, finančna, **migracijska, humanitarna,** ekonomska, **dolžniška,** bančna, politična, **begunsko-migrantska, mlečna, javnofinančna, varnostna,** kapitalistična kriza - [METAPHORICAL]: **socialna** kriza - [MODIFIERS]: huda, **kompleksna,** globoka, velika kriza - [NOUNS]: **začetek**, breme, **izbruh, nastop, posledica, nastanek,** reševanje, **obdobje** krize - [VERBS]: kriza **nastopi, nastane, pokaže, udari** // povzročiti, reševati, poglabljati krizo

As can be seen from Table 15, the biggest difference in relative frequency between the two corpora is observed for the noun *crisis*, which is more than twice as frequent in Parlameter compared to SlovParl, despite the fact that the early 1990s were marked by a long and bloody war in the Balkans as well as severe economic hardship related to change of the economic and political system. Parlameter contains the largest number of new collocation candidates that indicate issues that were not present in the period of SlovParl, such as *migrant/refugee/humanitarian/security crisis*. On the other hand,

the secession period was marked by *constitutional/parliamentary crisis*, which are not observed in the late 2010s. Interestingly, SlovParl contains more metaphorical collocations which are not prominent in the Parlameter corpus, such as *mental/social/welfare/moral crisis*. Collocations containing geographical terms indicate the key political, military and social hotspots from that period: *Yugoslav/Gulf crisis* in early 1990s, and *Ukraine/Greek crisis* in late 2010s. An analysis of key verbal collocates with the noun crisis reveals another interesting observation, which is that in SlovParl, all the verbs are about solving the crisis (*to solve/resolve/untangle the crisis*), whereas in Parlameter, politicians mostly use verbs that discuss the beginnings or deepening of the crisis (*crisis sets in/appears/starts/hits*, *to trigger/deepen the crisis*).

The verb *trigger* is the only one of the three examples that has a higher relative frequency in SlovParl but despite the greater relative frequency, Parlameter contains more collocation candidates, both in the direct and the metaphorical sense, such as *trigger an investigation/indictment/lawsuit*, or *trigger an audit/bankruptcy*.

It is interesting to note that the adjective *southern* is more frequently used and has more collocations in general in ParlaMeter despite the fact that in the secession period, links to the rest of former Yugoslavia were probably stronger and there were probably more open issues, signalling that certain topics were probably not discussed on purpose until the issues were resolved and the relations were established again. Especially interesting are all the neighbour-related collocations, which only appear in the Parlameter corpus, 30 years after Slovenia left Yugoslavia: *southern neighbour / neighbours / neighbourhood / market / fruit,* despite the fact that geographically speaking, the former Yugoslav republics, spread south-east, not south of Slovenia. The one major unsettled issue is the border with Croatia that has even been subject of international arbitration during the parliamentary term included in the Parlameter corpus, which is reflected in the top-ranking strong collocation *južna meja/southern border*.

Conclusions

In this paper we presented the Parlameter corpus of contemporary Slovene parliamentary proceedings. We analysed the linguistic production of the speakers according to the morphosyntactic annotation of the corpus and the speaker metadata.

We have shown that despite the fact that the material included in the corpus spans the period 2014–2018, the bulk of the material was recorded in the first two full years of the parliament. When contrasted against general Slovene, parliamentary speeches contain more present tense forms and personal and demonstrative pronouns. A comparison of male and female speakers shows that while male speakers take the floor more often than their female colleagues, it is the female speakers who make longer contributions. Female speakers mostly address the topics of *health, labour, family and social affairs, public administration,* and *education, science and sport*, while most of the keywords from male speakers do not belong to specific topics, which indicate a more

discursive, debating style of the male speakers. When comparing speeches according to party lines, the most prolific deputy group is the largest opposition party Slovenian Democratic Party (SDS) while the ruling Party of Modern Centre (SMC) is the least prolific one. The most productive parties with a relative token to speaker ratio are the smallest parties in this parliamentary term, the Left (Levica) and New Slovenia (NSi). The largest opposition party SDS stands out from the rest by the large amount of ideological keywords while Levica stands out by signature stylistic devices which range from very informal to highly elevated. NSi and Levica, the opposition parties with the same number of MPs but from the opposite ends of the political spectrum, both address the widest variety of issues. With keywords belonging almost exclusively to the semantic field of retirement and pension, DeSUS lies on the other end of the spectrum as a single-issue party. A comparison with the SlovParl corpus of parliamentary debates from the period of Slovenia's independence, many more topics are identified in Parlameter, which understandable as a well-established state will need to take care of a full spectrum of issues whereas a new state will mostly be dealing with procedural issues and the new legislature. In the future we plan to enrich the corpus with additional session records of previous and the most recent parliamentary terms as well as with additional metadata available through the Parlameter system, such as voting data and accepted legislation, which are also valuable for addressing a number of research questions in various research communities. In parallel, we also plan to develop comparable corpora from other parliaments, starting with Croatian and Bosnian.

Acknowledgments

The work described in this paper was funded by the Slovenian Research Agency within the national basic research project "Resources, methods, and tools for the understanding, identification, and classification of various forms of socially unacceptable discourse in the information society" (J7-8280, 2017–2019) and the Slovenian research infrastructure for language resources and technology CLARIN.SI.

Sources and Literature

Literature:

- Bayley, Paul. 2014. "Introduction: The Whys and Wherefores of Analyzing Parliamentary Discourse." In *Cross-Cultural Perspectives on Parliamentary Discourse*, edited by Paul Bayley, 1–44. Amsterdam, Philadelphia: John Benjamins Publishing.
- Cheng, Jennifer E. 2015. "Islamophobia, Muslimophobia or Racism? Parliamentary discourses on Islam and Muslims in Debates on the Minaret Ban in Switzerland." *Discourse & Society* 26 (5): 562–86.

- Chester, Daniel Norman, and Nona Bowring. 1962. *Questions in Parliament.* Oxford: Clarendon Press.
- van Dijk, Teun A. 2010. "Political Identities in Parliamentary Debates." In *European Parliaments Under Scrutiny: Discourse Strategies and Interaction Practices*, edited by Cornelia Ilie, 29–56. Amsterdam, Philadelphia: John Benjamins Publishing.
- Fišer, Darja, and Jakob Lenardič. 2018. "Parliamentary Corpora in the CLARIN Infrastructure." In *Selected Papers from the CLARIN Annual Conference 2017*, edited by Maciej Piasecki, 75–85. Accessed February 27, 2019. http://www.ep.liu.se/ecp/147/007/ecp17147007.pdf.
- Fišer, Darja, and Vojko Gorjanc. 2013. *Korpusna analiza*. Ljubljana: Znanstvena založba Filozofske Fakultete.
- Fišer, Darja, Nikola Ljubešić, and Tomaž Erjavec. 2018. "The Janes Project: Language Resources and Tools for Slovene user Generated Content." *Language Resources and Evaluation*. In press. https://doi.org/10.1007/s10579-018-9425-z.
- Franklin, Mark N., and Philip Norton. 1993. *Parliamentary Questions: For the Study of Parliament Group*. Oxford: Oxford University Press.
- Hirst, Graeme, Vanessa Wei Feng, Christopher Cochrane, and Nona Naderi. 2014. "Argumentation, Ideology, and Issue Framing in Parliamentary Discourse." In *ArgNLP*. Accessed 27 February 2019. ftp://www.cs.toronto.edu/pub/gh/Hirst-etal-Bertinoro-2014.pdf.
- Hughes, Lorna M., Paul S. Ell, Gareth A.G. Knight, and Milena Dobreva. 2013. "Assessing and Measuring Impact of a Digital Collection in the Humanities: An Analysis of the SPHERE (Stormont Parliamentary Hansards: Embedded in Research and Education) Project." *Digital Scholarship in the Humanities* 30 (2): 183–98.
- Ihalainen, Pasi, Cornelia Ilie, and Kari Palonen. 2016. *Parliament and Parliamentarism: A Comparative History of a European Concept.* Oxford, New York: Berghahn Books.
- Ilie, Cornelia. 2017. "Parliamentary Debates." In *The Routledge Handbook of Language and Politics*, edited by Ruth Wodak and Bernhard Forchtner. Routledge.
- Ljubešić, Nikola, and Tomaž Erjavec. 2016. "Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: The Case of Slovene." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 1527–31. Accessed February 27, 2019. http://www.lrec-conf.org/proceedings/lrec2016/pdf/811_Paper.pdf.
- Ljubešić, Nikola, Tomaž Erjavec, Darja Fišer, Tanja Samardžić, Maja Miličević, Filip Klubička, and Filip Petkovski. 2016. "Easily Accessible Language Technologies for Slovene, Croatian and Serbian." In *Proceedings of the Conference on Language Technologies and Digital Humanities 2016*, edited by Tomaž Erjavec and Darja Fišer, 120–24. Accessed February 27, 2019. http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Ljubesic-et-al_Easily-Accessible-Language-Technologies.pdf.
- Ljubešić, Nikola, Darja Fišer, Tomaž Erjavec, and Filip Dobranić. 2018. "The Parlameter corpus of contemporary Slovene parliamentary proceedings." In *Proceedings of the Conference on Language Technologies and Digital Humanities 2018,* edited by Darja Fišer and Andrej Pančur, 162–167. Accessed June 12, 2019. http://www.sdjt.si/wp/wp-content/uploads/2018/09/JTDH-2018_Ljubesic-et-al_The-Parlameter-corpus-of-contemporary-Slovene-parliamentary-proceedings.pdf
- Pančur, Andrej, and Mojca Šorn. 2016. "Smart Big Data: Use of Slovenian Parliamentary Papers in Digital History." *Prispevki za novejšo zgodovino* 56 (3): 130–46.
- Pančur, Andrej. 2016. "Označevanje zbirke zapisnikov sej slovenskega parlamenta s smernicami TEI." In *Proceedings of the Conference on Language Technologies and Digital Humanities 2016*, edited by Tomaž Erjavec and Darja Fišer, 142–48. Accessed February 27, 2019. http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Pancur_Oznacevanje-zbirke-zapisnikov-sej-slovenskega-parlamenta.pdf.

- Rheault, Ludovic, Kaspar Beelen, Christopher Cochrane, and Graeme Hirst. 2016. "Measuring Emotion in Parliamentary Debates with Automated Textual Analysis." *PLoS ONE* 11 (12): 1–18.
- TEI Consortium, 2017. Guidelines for Electronic Text Encoding and Interchange. Accessed February 27, 2019. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html.

Sources:

- Dobranić, Filip, Nikola Ljubešić, and Tomaž, Erjavec. 2019. *Slovenian Parliamentary Corpus ParlaMeter-sl 1.0*, Slovenian Language Resource Repository CLARIN.SI. http://hdl.handle.net/11356/1208.
- Pančur, Andrej, Mojca Šorn, and Tomaž Erjavec. 2017. *Slovenian Parliamentary Corpus SlovParl 2.0*, Slovenian Language Resource Repository CLARIN.SI. http://hdl.handle.net/11356/1167.

Darja Fišer, Nikola Ljubešić, Tomaž Erjavec

PARLAMETER – A CORPUS OF CONTEMPORARY SLOVENE PARLIAMENTARY PROCEEDINGS

SUMMARY

The unique content, structure and language, as well as the availability of records of parliamentary debates are all factors that make them an important object of study in a wide range disciplines in digital humanities and social sciences. This has motivated a number of national as well as international initiatives to compile, process and analyse parliamentary corpora. This paper presents the Parlameter corpus of contemporary Slovene parliamentary proceedings, which covers the VIIth mandate of the Slovene Parliament (2014–2018). The Parlameter corpus offers rich speaker metadata (gender, age, education, party affiliation) and is linguistically annotated (lemmatization, tagging, named entity recognition).

The Parlameter corpus contains 371 sessions and 1,981 speakers who gave 133,287 speeches which contain almost 35 million words. In the paper we demonstrate the potential of the corpus analysis techniques for investigating political debates by analysing the linguistic production of the speakers according to the morphosyntactic annotation of the corpus and the speaker metadata. When contrasted against general Slovene, parliamentary speeches contain more present tense forms and personal and demonstrative pronouns. While male speakers take the floor more often than their female colleagues, the female speakers' contributions tend to be longer. Female speakers mostly address the topics of health, labour, family and social affairs, public administration, and education, science and sport, while most of the keywords from male speakers do not belong to specific topics, which indicate a more discursive, debating style of the male speakers. The most prolific deputy group overall is

the largest opposition party Slovenian Democratic Party (SDS) while the then ruling Party of Modern Centre (SMC) is the least prolific. The most productive parties with a relative token to speaker ratio are the smallest parties in that parliamentary term, the Left (Levica) and New Slovenia (NSi). The largest opposition party SDS stands out from the rest by the large amount of ideological keywords while Levica stands out by signature stylistic devices which range from very informal to highly elevated. NSi and Levica, the opposition parties with the same number of MPs but from the opposite ends of the political spectrum, both address the widest variety of issues. With keywords belonging almost exclusively to the semantic field of retirement and pension, DeSUS lies on the other end of the spectrum as a single-issue party. A comparison with the SlovParl corpus of parliamentary debates from the period of Slovenia's independence, many more topics are identified in Parlameter, which understandable as a well-established state will need to take care of a full spectrum of issues whereas a new state will mostly be dealing with procedural issues and the new legislature.

The Parlameter corpus is available through both CLARIN.SI concordancers, as well as for download from its repository, both as a TEI document and in the simpler vertical file format, under the liberal Creative Commons – Attribution-ShareAlike (CC BY-SA 4.0) licence. The corpus architecture allows for regular extensions of the corpus with additional Slovene data, as well as data from other parliaments, starting with Croatian and Bosnian.

Darja Fišer, Nikola Ljubešić, Tomaž Erjavec

PARLAMETER – KORPUS RAZPRAV SLOVENSKEGA DRŽAVNEGA ZBORA

POVZETEK

Edinstvena vsebina, struktura in jezik, pa tudi dostopnost prepisov parlamentarnih razprav so dejavniki, zaradi katerih so le-ti pomemben predmet raziskav v številnih znanstvenih disciplinah digitalne humanistike in družboslovja. To je motiviralo številne nacionalne in mednarodne iniciative za izgradnjo, označevanje in analizo parlamentarnih korpusov. V tem prispevku predstavimo korpus sodobnih parlamentarnih razprav Parlameter, ki vsebuje razprave 7. mandata slovenskega Državnega zbora (2014–2018). Korpus Parlameter vsebuje bogate metapodatke o govorcih (spol, starost, izobrazba, strankarska pripadnost) in je jezikoslovno označen (lematizacija, tegiranje, imenske entitete).

Korpus Parlameter vsebuje 371 razprav in 1.981 govorcev, ki so prispevali 133.287 govorov oziroma 35 milijonov besed. V prispevku prikažemo potencial korpusno-analitičnih tehnik za raziskovanje političnih razprav z analizo jezikovne produkcije

govorcev glede na morfosintaktične oznake in metapodatke o govorcih. Primerjava s splošno slovenščino pokaže, da v parlamentarnih govorih izstopajo sedanjiške oblike ter osebni in kazalni zaimki. Čeprav moški govorci spregovorijo večkrat, so govori žensk daljši. Ženske večinoma razpravljajo o temah, kot so zdravje, delo, družina in sociala, javna uprava ter izobraževanje, znanost in šoprt, večina ključnih besed v moških govorih pa ni vezanih na določeno tematiko, kar nakazuje bolj diskurziven, razpravljalski slog moških govorcev. V celoti gledano je najbolj produktivna strankarska skupina največja opozicijska stranka SDS, medtem ko je vladajoča stranka SMC v korpusu zastopana z najmanj izrečenimi besedami. Najvišji relativni delež števila pojavnic na govorca imata najmanjši parlamentarni stranki tega sklica Levica in NSi. Največja opozicijska stranka SDS izstopa po izrazito velikem obsegu ideološko obarvanih ključnih besed, Levica pa po specifičnih slogovnih figurah, ki so tako zelo neformalne kot zelo povzdignjene. NSi in Levica, opozicijski stranki z enakim številom poslancev a s povsem različnih polov političnega spektra, obe naslavljajta največje število tematik. Po drugi strani pa s ključnimi besedami, ki skoraj v celoti spadajo v pomensko polje upokojevanja in pokojnin, pa je povsem obratno pri stranki DeSUS, ki s tem utrjuje svoj status problemske stranke. Primerjava s korpusom SlovParl iz obdobja slovenske osamosvojitve kaže, da je v korpusu Parlameter obravnavanih veliko več tem kot v korpusu SlovParl, kar je razumljivo, saj se mora uveljavljena država ukvarjati s celotnim spektrom problematik, medtem ko se novo ustanovljena država posveča predvsem priceduralnim vprašanjem in sprejemanju nove zakonodaje.

Korpus Parlameter je dostopen preko obeh konkordančnikov v okviru raziskovalne infrastructure CLARIN.SI, prav tako pa ga je mogoče prenesti z repozitorija v format TEI, pa tudi v preprostejšem vertikalnem formatu pod licenco Creative Commons – Attribution-ShareAlike (CC BY-SA 4.0). Korpusna arhitektura je zasnovana tako, da omogoča širitev korpusa na druga časovna obdobja, prav tako pa tudi vključevanje gradiv drugih parlamentov, začenši s hrvaškim in bosanskim.

1.01 UDC: 003.295:821.163.6'367.625

Polona Gantar,[*] Špela Arhar Holdt,[**] Jaka Čibej,[***] Taja Kuzman[****]

Structural and Semantic Classification of Verbal Multi-Word Expressions in Slovene

IZVLEČEK

STRUKTURNA IN POMENSKA KLASIFIKACIJA GLAGOLSKIH VEČBESEDNIH ENOT V SLOVENŠČINI

Prispevek je nadgrajena različica konferenčnega prispevka, v katerem predstavljamo kategorije glagolskih večbesednih enot (GVBE), kot so bile oblikovane v okviru mednarodne COST akcije PARSEME Shared Task 1.1. S kategorijami, ki so nadjezikovne in obenem prilagojene posameznim vključenim jezikom, smo označili 13.511 povedi učnega korpusa ssj500k 2.0. Rezultat označevanja je 3.364 identificiranih večbesednih glagolskih enot, ki so klasificirane kot: inherentno povratni glagoli, zveze z glagoli v pomensko oslabljeni rabi, predložnomorfemski glagoli in glagolski idiomi. V prispevku rezultate označevanja predstavimo kvantitativno in kvalitativno, pri čemer sopostavimo predlagani sistem klasifikacije ob obstoječe prakse na področju slovenistične obravnave GVBE in ocenimo uporabnost sistema za nadaljnje delo.

Ključne besede: glagolske zveze, korpusni pristop, večbesedne enote, PARSEME, slovenščina

[*] **Department of Translation, Faculty of Arts, University of Ljubljana, Aškerčeva 2, SI-1000 Ljubljana, apolonija.gantar@guest.arnes.si**

[**] **CJVT, Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, SI-1000 Ljubljana, spela.arhar@cjvt.si**

[***] **Artificial Intelligence Laboratory, Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, jaka.cibej@ijs.si**

[****] **kuzman.taja@gmail.com**

ABSTRACT

This paper is an extended version of a conference paper presenting the categorization of verbal multi-word expressions (VMWEs) according to the PARSEME COST Action Shared Task 1.1 Guidelines. The categorization is universal but takes into account the characteristics of the individual languages included in it. The Shared Task was used to annotate over 13,000 sentences of the Slovene ssj500k 2.0 training corpus, which resulted in nearly 3,400 identified VMWEs categorized as inherently reflexive verbs, light verb constructions, inherently adpositional verbs, and verbal idioms. The paper presents both the quantitative and qualitative results of the analysis, compares the suggested categorization system to existing work on VMWEs in Slovene linguistics, and evaluates the use of the proposed system for future work.

Keywords: verb phrases, corpus approach, multi-word expressions, PARSEME, Slovene

Introduction

In the digital medium, the bulk of interactions between users – as well as between users and computers or applications – occur with the use of language, which is why the existence and open accessibility of digital language infrastructure is of vital importance to the development and vitality of a language. Slovene is no exception in this regard; it requires an infrastructure that serves as a source of information for the language community as well as a resource to be used in applied/theoretical linguistic research and the development of new language technology tools, methods, and services. Examples of such infrastructure include digital language resources that allow for continued updates and contributions from the community, language databases with structured and machine-readable data, and training corpora in which authentic texts are annotated with different linguistic categories. In this regard, digital lexicography, whose aim is to prepare the dictionary part of this language infrastructure, plays an essential role in the field of digital humanities.

In the field of digital lexicography, multi-word expressions (MWEs) are considered important for constructing machine-readable language resources that in turn enable the compilation of electronic MWE lexicons and the development of language technology tools for a specific language. In order to achieve these goals, it is crucial to know the linguistic features of different types of MWEs and develop methods and standards for their identification in authentic language use.

However, this is not a trivial task. Definitions and categorisations of MWEs differ according to their methodological and theoretical basis and research goals.[1] A lexicographic perspective focuses on the semantic characteristics of MWEs and defines

1 For an overview of MWE classifications according to different methodological approaches, see Gantar et al. (2018).

them as "different types of phrases that demonstrate a certain degree of idiomatic meaning" (Atkins and Rundell 2008, 166) or as phrases whose "exact meaning is not directly obtained from its component parts" (Sag et al. 2002). On the other hand, the definition of MWEs from the perspective of machine processing emphasises their statistical significance: "a group of tokens in a sentence that cohere more strongly than ordinary syntactic combinations. That is, they are idiosyncratic in form, function, or frequency" (Schneider et al. 2014) and their inability to be split into independent lexemes and at the same time maintain their semantic and syntactic functions, as well as their lexical, syntactic, semantic, pragmatic and statistical idiomaticity (Baldwin and Kim 2010, 3). Although no universally accepted definition of MWEs exists, researchers in linguistics and NLP both agree that the key feature separating MWEs from free phrases is the special relationship between the elements that form the MWE. This relation is usually defined using such concepts as collocability (or statistical idiomaticity), idiomaticity (or semantic non-compositionality), syntactic (in)flexibility, which also includes the possibility of an internal modification of the phrase and the flexible order of its lexicalised elements, and lexical variance.

An attempt to provide the much needed guidelines and a pilot study on the annotation of MWEs in language corpora was made as part of the PARSEME COST Action Shared Task 1.1.[2] The task focused on the automatic identification of verbal multi-word expressions (VMWEs) in running text. As part of the task, universal guidelines for VMWE classification were compiled containing examples for all languages involved. Based on the guidelines, a multi-lingual corpus was manually annotated with VMWEs and made available under a Creative Commons licence.

While the categories of MWEs were designed as language-independent, the specific characteristics of all the included languages had to be taken into account to reach a universally applicable solution. In this paper, we focus on the Slovene results, which will be useful when compiling a digital lexicon of Slovene MWEs, as well as other language resources such as the Dictionary of Modern Slovene (Gorjanc et al. 2017) and a corpus-based grammar of Slovene. The topic was presented in Gantar et al. (2018) with a focus on MWEs and their theoretical framework in Slovene studies. This paper focuses on MWEs from the perspective of a unified concept that was applied to 20 different languages within the PARSEME Shared Task 1.1. A comparison of the results can be found in Ramisch et al. (2018).

Identifying and Categorizing Verbal Multi-Word Expressions

The verb plays a crucial role in the sentence in terms of co-text organization, which is why the PARSEME Shared Task focused on verbal multi-word expressions

2 *Home – PARSEME*, http://www.parseme.eu.

(VMWEs). For further analysis, it is crucial to determine the differences between the definitions and categorizations of VMWEs as established in Slovene studies on the one hand, and the international PARSEME COST Action on the other. Our task aims to adopt the international annotation scheme in order to include Slovene. Our research question focuses on the applicability of the PARSEME system to authentic Slovene texts. Can the adapted PARSEME categories be applied in practice? Are they attributable, robust, one-dimensional, and represented in actual language use? What information do they entail (e.g. in terms of syntax), how can they contribute to the development of new automatic extraction methods, and finally, which problems arise when applying the system to text? In the following sections, we present the annotation method. This is followed by quantitative and qualitative analysis. The latter is focusing on individual categories, their characteristics, and the potential problems of the approach.

Verbal Multiword Expressions – Slovenian Case

In Slovene studies, MWEs are divided into a) phraseological units (PUs), in which at least one component carries meaning that differs from one of its denotative "dictionary" senses, and expresses figurativeness, and b) all other multiword expressions (i.e. fixed expressions), which are characterized by a certain degree of fixedness and denote a meaning that can be predicted from the meanings of their elements. PUs are further divided by syntactic structure: the clausal type (which also includes proverbs) and the phrasal type (all non-verbal PUs). In Slovene linguistic theory, verbal MWEs are determined by their morphosyntactic features (Toporišič 1973/74; Kržišnik 1994): a MWE is classified as a VMWE if it includes a verbal element and if it functions as a predicate. However, it remains unclear how to classify examples in which the verbal MWE does not function as a predicate, e.g. *hočeš nočeš* 'like it or not', which includes two verbal elements, but functions as an adverbial.

The problem of categorizing MWEs according to their morphological structure and syntactic function was resolved in PARSEME shared task through the definition that the main criterion for VMWEs is that their syntactic head in the prototypical form is a verb, regardless of the fact whether it can or cannot fulfil other syntactic roles. In addition, Slovene categorizations have so far never treated verbs with the *se/si* morpheme as a separate MWE category. Phrasal verbs that consist of a verb and a preposition and carry an independent meaning were categorized as MWEs only conditionally (Kržišnik 1994, 58).

Verbal Multiword Expressions within the Parseme Shared Task 1.1

For the categorization of VMWEs within the Parseme Shared task 1.1, exhaustive guidelines[3] were prepared in which the VMWE categories are defined by semantic and syntactic features and are described with decision trees. The identification and categorization process consisted of three steps. In the first step, we identified potential VMWEs consisting of a verb as the syntactic head of the phrase and at least one other word. In the second step, we identified the lexicalised elements within the phrase. In the third step, we used detailed linguistic tests consisting of generic and specific language criteria to determine the correct category of the identified VMWE.

Based on the guidelines, VMWEs are further divided into two classes based on whether the category can be applied to the majority of languages included in the task, or whether they are typical of individual (groups of) languages. The universal categories include verbal idioms (VID) and light verb constructions (LVC), which are further divided into full (LVC.full) and causal (LVC.cause). The quasi-universal categories, which are used within individual groups of languages, include inherently reflexive verbs (IRV), which are typical of most Slavic languages, and verb-particle constructions (VPC), typical of Germanic languages. In the second version of the guidelines, an additional quasi-universal category was added: inherently adpositional verbs (IAV), which require an open syntactic slot and are typical of Slovene and several other Slavic languages.

For Slovene, examples of VMWEs can be found for all the categories except for VPC. For certain categories, however, there are specific characteristics based on syntactic or morphological features of Slovene or on grammatical categories that are generally accepted in Slovene but differ to some extent from other languages. The specific Slovene features will be described along with individual VMWE types.

The Corpus and Annotation Tool

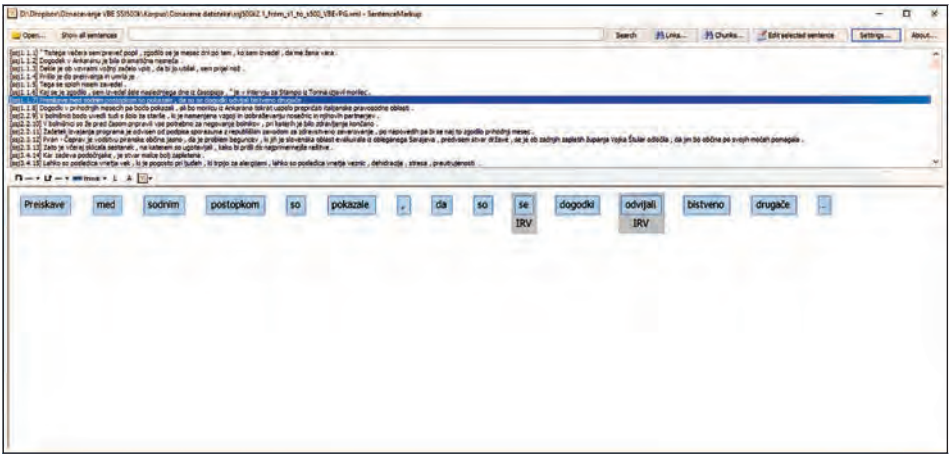
VMWEs were annotated in the Slovene ssj500k 2.0 training corpus (Krek et al. 2017), which consists of approximately 500,000 tokens and just under 28,000 sentences sampled from the FidaPLUS corpus of Slovene (Arhar Holdt and Gorjanc 2007). The entire corpus is morphosyntactically tagged (Grčar et al. 2012). Certain portions also contain named-entity annotations and syntactic dependencies (Dobrovoljc et al. 2012). In the first annotation phase, 11,411 sentences were annotated by two annotators with VMWEs as defined by the first version of the PARSEME Guidelines (Candito et al. 2016). Disagreements in annotation were discussed and adjusted accordingly. In the second phase, the categories were automatically modified based on the second version of the PARSEME Guidelines and manually checked. The

3 *PARSEME Shared Task 1.1 - Annotation guidelines*, http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/.

second phase continued with the annotation of an additional 2,100 sentences annotated in packages by individual annotators. Problematic examples were discussed and correctly annotated.

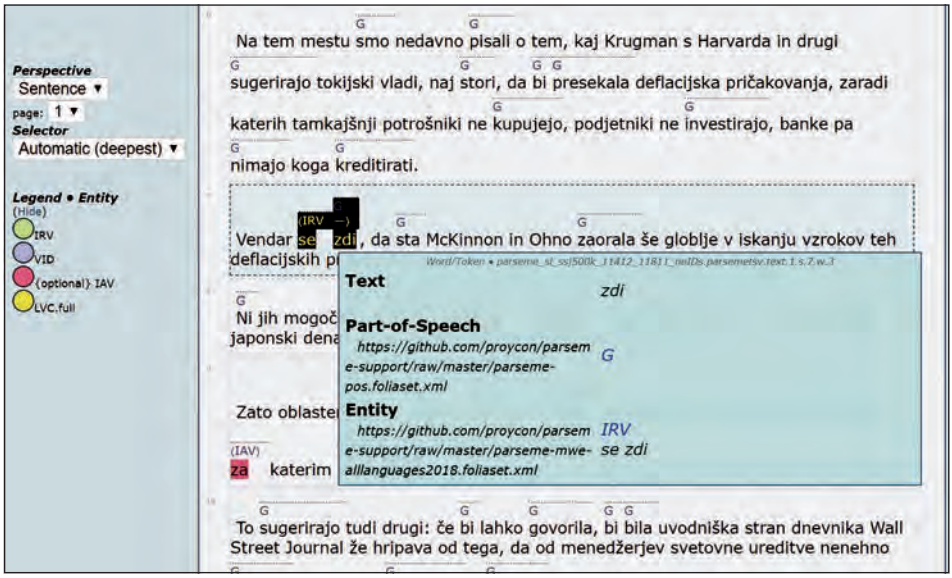
The tool used for annotation in the first phase was SentenceMarkup System (Figure 1), a custom tool primarily developed for syntactic dependency annotation of Slovene texts (Dobrovoljc et al. 2012). The tool was adjusted for the annotation of VMWEs by adding an additional independent and interconnectable annotation layer (cf. Gantar et al. 2017).

Figure 1: Annotations in the SentenceMarkup System



In the second phase, the annotation took place in the FLAT annotation platform (FoLiA Linguistic Annotation Tool), which was adapted for the purposes of the PARSEME Shared Task and tested on 13 collaborating languages (Figure 2). The FLAT platform allows text strings to be annotated with a set of categories. Files can be assigned to different annotators. The supported formats are XML and TSV, while annotated files are exported in XML. All annotations are saved automatically. The interface also features text search using CQL.

Figure 2: Annotations in FLAT



Quantitative Analysis

The annotated VMWEs were imported into the ssj500k 2.1 training corpus (Krek et al. 2017). Among the 13,511 sentences annotated in the first two annotation phases, 2,290 of them (approximately 22%) contain at least one VMWE. On average, each of these sentences features 1.15 VMWEs. Taking into account all the annotated sentences, each sentence contains approximately 0.25 VMWEs; in other words, on average, one VMWE is present in every fourth sentence.

Table 1 shows the distribution of the annotated VMWEs by category. The final number of VMWEs in the training corpus is 3,364. The number of different VMWEs (i.e. without any repetitions of the same unit) was just under 1,100. When looking at absolute frequencies, the most frequent category is IRV (48%) and the least frequent category is LVC.cause (2%). The categories with the highest number of different VMWEs are VID and IAV, while LVC.full and LVC.cause are the least diverse categories. We describe each category in more detail in section 5.

Table 1: Distribution of annotated VMWEs by category

Category	Example	Translation	All VMWEs	Percent	Different VMWEs
Inherently Reflexive Verbs (IRV)	*bati se*	to be afraid	1,627	48%	345
Inherently Adpositional Verbs (IAV)	*priti do*	to come about	710	21%	154
Verbal Idioms (VID)	*spati kot ubit*	(lit.) to sleep like a dead person	724	22%	457
Light Verb Constructions (LVC): LVC.cause	*spraviti koga v smeh*	to make someone laugh	64	2%	27
Light Verb Constructions (LVC): LVC.full	*biti v pomoč*	to be of help	239	7%	103
Total	-	-	3,364	100%	1,086

Table 2 shows the most common VMWE structures by parts of speech (V – verb, N – noun, A – adjective, R – adverb, Pre – preposition, Pro – pronoun). The structures occurring in the corpus with a frequency below 10 have been categorized as Other. The most frequent structures are V + Pro, V + Pre, V + N and V + Pre + N. Collectively, they account for approximately 85% of all annotated VMWEs.

Table 2: Distribution of annotated VMWEs by part-of-speech structure

Structure	Example	Translation	Frequency	Percent
V + Pro	*bati se*	to be afraid	1,663	49%
V + Pre	*priti do*	to come about	535	16%
V + N	*imeti odnos*	to have a relationship	372	11%
V + Pre + N	*biti pod vtisom*	to be under the impression	303	9%
V + Pro + A	*biti si edini*	to be unanimous	146	4%
V + R	*biti res*	to be true	136	4%
V + Pro + Pre + N	*ujeti se v past*	to get caught in a trap	24	1%
V + A	*biti jasno*	to be clear	20	1%
V + A + N	*imeti glavno besedo*	to have the last word	19	1%
N + V + Pre + N	*biti na robu propada*	to be on the verge of collapse	12	<1%
V + Pro + N	*vzeti si čas*	to take one's time	11	<1%
Other	-	-	123	4%
Total	-	-	3,364	100%

Qualitative Analysis

The qualitative analysis deals with the semantic and structural features of VMWEs. Based on the PARSEME Guidelines, several characteristic features of Slovene were identified on the level of structural and semantic tests used to determine the category of VMWEs. In the analysis, we focused on patterns within structures for each subcategory, the syntactic environment of the expression as a unit, and the lexical units filling the corresponding participant slots. Based on corpus examples, we also tried to identify the indicators of semantic integrality that could be useful when automatically identifying VMWEs in text.

Inherently Reflexive Verbs (IRV)

The PARSEME Shared Task 1.1 guidelines treat verbs occurring with the independent morpheme *se/si* as a separate category of VMWEs called *inherently reflexive verbs*. It is a language-specific category that includes phrases in which the verb without the morpheme *se/si* does not exist (*zdeti se* 'to seem', **zdeti*) or in which the presence of *se/si* changes the meaning of the verb (*pobrati se* 'to recover' vs. *pobrati* 'to pick up').

Inherently reflexive verbs cover the largest percentage of VMWEs in the training corpus (see Table 1). Among the correctly categorized examples (1,621 in total)[4] we identified 339 different IRVs, with the following most frequently occurring verbs: *zdeti se* 'to seem', *odločiti se* 'to decide', *zgoditi se* 'to come to pass' and *pojaviti se* 'to appear'.

To test whether the expression is semantically integral and to differentiate it from other types of verb phrases with *se/si* that are not defined as VMWEs, we examined the behaviour of the verb in terms of its opening up syntactic positions as a phrase. Inherently reflexive verbs keep *se/si* as an obligatory verb morpheme in all forms of their inflectional paradigm and can be transitive (*bati se koga/česa* 'to be afraid of smn/sth') or intransitive (*znajti se* 'to find oneself somewhere', *zvečeriti se* 'to fall [evening]').

Inherently reflexive verbs as VMWEs must be differentiated from verbs where the reflexive pronoun *se/si* is not an obligatory morpheme but serves another function, more specifically: (a) it denotes mutualness (*poljubljati se* 'to kiss [each other]', *srečati se* 'to encounter [each other]'), (b) it denotes that the target of the action is the subject (*umivati se* 'to wash [oneself]', *praskati se* 'to scratch [oneself]'), or that the action is to the benefit of the subject (*kuhati si* 'to cook [oneself sth]', (c) it is used for passivizing the sentence by removing the agent (*kdo ponavlja kaj* 'someone repeats something' – *kaj se ponavlja* 'something is repeated'), and (d) it denotes a generic action (*govori se* 'it is said'; *se razume* 'it is understood').

With verbs that can also occur without *se/si*, only the phrases where the morpheme changes the verb's meaning are categorized as IRVs. There are cases in which

4 Among the 1,627 annotated examples, four were miscategorized. In two examples, the elements of the expressions were incorrectly annotated. These examples were excluded from further analysis.

the presence (or absence) of *se/si* causes a semantic shift directly tied to a human subject In these cases, the verb denotes a metaphorical meaning *pobrati se* 'to recover': *pobrati* 'to pick sth up'; *gristi se* 'to worry' : *gristi* 'to bite'.

In Slovene linguistics, lexicalised phrases consisting of a verb and the *se/si* morpheme have so far not been treated as fixed expressions. The main focus has been recognition of the function of the morpheme or the reflexive pronoun in terms of denoting different degrees of agentness or the subject's (un)involvedness, as in the case of the non-singular (*zbrati se* 'to gather') or generic agent (*tiskati se* 'to be printed') (Žele 2012, 44; Toporišič 1982, 244; 2000, 503). The identification of IRVs in text from the perspective of their semantic and syntactic is particularly important for the automatic identification of MWEs. In future lexicons and dictionaries, IRVs should thus be treated either as independent entries or as part of polysemy.

Light Verb Constructions (LVC)

Light verb constructions have been treated from different perspectives by different authors (for an overview, see Soršak 2013). In most definitions, the verbs in LVCs are categorized as something between full verbs and auxiliary verbs, while the expressions that feature them are categorized as a phenomenon between fixed and free expressions. Using existing typologies for Slovene (Toporišič 2000; Žele 1999), Soršak analyzes Slovene LVCs based on the entries in the Dictionary of Standard Slovene (SSKJ). The results highlight that the dictionary often mentions the semantically light use of a verb in places where the use is stylistically marked, most frequently as expressive (Soršak 2013, 514; e.g. *groza ga sprehaja,* lit. 'terror is walking him'). The results described in this paper show the opposite – in the annotated corpus, LVCs are typical, stylistically neutral, and frequently occurring.

As per the PARSEME Guidelines, a LVC must fulfil the following conditions: it consists of a verb and a noun or a noun phrase that can also take the form of a prepositional phrase (*imeti mnenje* 'to have an opinion', *biti v dvomih* 'to be in doubt'), and must open up its own valency slots (*kdo ima predavanje za koga* 'someone holds a lecture for someone'). Semantically, the expression must denote an action (*imeti predavanje* 'to hold a lecture') or a state (*biti v dvomih* 'to be in doubt'). According to the verb, the category has two subtypes: (a) if the verb contributes to the meaning on a predominantly categorical level, the expression is categorized as LVC.full (*biti v pomoč* 'to be of help'); (b) if the subject can be interpreted as the cause or source of the denoted action, the expression is categorized as LVC.cause (*spraviti v smeh* 'to make smn laugh'). The LVC tests also take into account the abstractness of the noun (*imeti avto* 'to have a car' is not a multiword expression, while idiomatic expressions like *imeti mačka* 'lit. to have a cat – to have a hangover' are categorized as VIDs) and, with LVC.full, the possibility of rephrasing by omitting the verb (*Janez ima predavanje* 'Janez holds a lecture' – *Janezovo predavanje* 'Janez's lecture').

Despite the somewhat elusive concept of LVCs, the annotation process has confirmed that the PARSEME guidelines are specific enough to be successfully applied to real text. Of the 303 examples annotated as LVCs (1 example was categorized incorrectly), 78.8% were LVC.full and 21.2% LVC.cause. 87.1% of them were combinations of a verb and a noun, while 12.9% were combinations of a verb and a prepositional phrase. The annotated LVCs contained a total of 19 different verbs,[5] predominantly the verb *imeti* 'to have' (65.6%), but also *biti* 'to be' (13.6%) and *dati/dajati* 'to give' (a total of 9.6%).[6] Other verbs (*narediti* 'to do', *postaviti/postavljati* 'to put', *ostati* 'to remain', *voditi* 'to lead', *namenjati* 'to pay [attention]', *delati* 'to do/make', *storiti* 'to do', *vzbujati/zbujati* 'to incite', *dobiti* 'to get', *zastaviti* 'to pose', *spraviti* 'to make', *doseči* 'to achieve' and *nositi* 'to wear') occur less frequently, often in a single expression (*ostati v spominu* 'to remain in one's memory', *namenjati pozornost* 'to pay attention to sth').

Combinations of a verb and a prepositional phrase are somewhat more typical of the LVC.cause category. In the annotated data, LVC.cause occurs exclusively with the prepositions *v* 'in' (33 instances) and *na* 'on' (6 instances). In the majority of cases, the combination is *biti v* (*biti v pomoč* 'to be of help', *biti v podporo* 'to provide support', *biti v navadi* 'to be a habit').

In the annotated expressions, a relatively limited set of nouns can be found: a total of 97. The most frequent nouns are *težava* 'problem' (21) and *pravica* 'right' (20), followed by *možnost* 'possibility', *mnenje* 'opinion', *učinek* 'effect', *vloga* 'role', *vpliv* 'influence', *vtis* 'impression', *pomoč* 'help', *občutek* 'feeling', *prednost* 'advantage', *sreča* 'luck', *korist* 'benefit', *vprašanje* 'question', *volja* 'will', *posledica* 'consequence'. As expected, some of these nouns occur exclusively in LVC.full (*pravica, možnost, mnenje, vloga*), while others occur in LVC.cause (*učinek, vpliv, vtis, pomoč*). In other cases, the category depends on the meaning of the verb (*dati prednost* 'to give an advantage' ® LVC.cause and *imeti prednost* 'to have an advantage' ® LVC.full.

In accordance with the conclusions made by Soršak (2013, 519), the results show that the featured verbs can also be used with full meaning, while the semantic lightness in LVCs is complemented by the nominal part (*imeti* 'to have' meaning 'to possess' compared to *imeti posledice* 'to have consequences' meaning 'to cause/lead to consequences'). Semantically, the noun groups occurring in LVC.cause describe the result of an action, be it a type of result (*učinek* 'effect', *vpliv* 'influence', *vtis* 'impression'), a positive (*korist* 'benefit', *užitek* 'pleasure') or negative consequence (*muka* 'torment', *preglavica* 'trouble'). The semantically light verb binds the result to the subject (*nekdo/nekaj daje vtis* 'smn/sth makes an impression', i.e. the agent is the cause of the action). In certain cases, LVCs can be converted into semantically full verbs with a similar morphological base (*dosegati učinek* 'to achieve an effect' – *učinkovati* 'to affect'; *imeti*

5 This is the full set of the LVCs in the data, confirming that the set of verbs occurring in these expressions is limited. In the dictionary, Soršak (2013, 513) finds mentions of semantic lightness in 420 verb entries. However, as mentioned, the labels often signify stylistically marked and atypical language use.

6 In Slovene lingustics, verb phrases with *imeti* 'to have' and *biti* 'to be' have been most frequently treated as the equivalent of LVCs, but analyzed from different perspectives (see e.g. Vidovič Muha 1998).

vpliv 'to have an influence' – *vplivati* 'to influence'), but not always (*imeti posledice* 'to have consequences' – /).

The nouns occurring in LVC.full are semantically more diverse. Dividing them into semantic groups reveals that the common ground of these expressions can be defined as planning or estimating success. Among the encountered LVCs are phrases with nouns dealing with (a) communication (*mnenje* 'opinion', *predlog* 'suggestion', *vprašanje* 'question'); or describing (b) the potential for success (*možnost* 'possibility', *prednost* 'advantage', *priložnost* 'opportunity'); (c) initial steps (*obljuba* 'promise', *napoved* 'prediction', *načrt* 'plan'); (d) potential reasons for failure (*napaka* 'mistake', *pomanjkljivost* 'disadvantage'). Other groups deal with (e) negative states (*težava* 'problem', *strah* 'fear', *dvom* 'doubt'), (f) positive qualities (*moč* 'power', *pogum* 'courage', *potrpljenje* 'patience'), (g) achieved results (*izobrazba* 'education', *status* 'status', *posel* 'business'), and (h) attitude towards as of yet unrealized goals (*želja* 'wish', *ambicija* 'ambition', *vizija* 'vision'). Again, some examples can be converted into a semantically full verb (*imeti mnenje* 'to have an opinion' – *meniti*), while others cannot (*imeti ambicije* 'to have ambitions' – /).

Inherently Adpositional Verbs (IAV)

Inherently adpositional verbs, also called verbs with a lexicalised prepositional morpheme (Žele 2002), were included in the PARSEME Guidelines during the second annotation phase as an optional test category.[7] The guidelines define IAVs as verbs that only occur with a prepositional morpheme (*simpatizirati z* 'to sympathize with') or verbs that change meaning when occurring with a prepositional morpheme (*biti za* 'be for, to support' vs. *biti* 'to be'). The participants required by the verb phrase as a whole are not a part of the VMWE, as opposed to e.g. *stati na + trdnih tleh* 'to stand on + solid ground', which is categorized as a VID.

Prepositions have been treated as free verb morphemes as early as in Metelko's Grammar of Slovene (1825, 247–56) and were analyzed in further detail by Breznik (1916, 250; 1934, 225). Verbs with a lexicalised prepositional morpheme were also analyzed by Žele (2002) and Kržišnik (1994), the former from the perspective of the degree of lexicality of the preposition and the latter from the perspective of phrase fixedness as either a phraseological unit with structural fixedness (*biti ob čem* 'to be next to sth' meaning 'to be positioned next to sth') or phrasemes with lexical fixedness (*biti ob kaj* 'to lose sth').

7 Based on the feedback from the first annotation campaign and the issues discussed among the contributors, idiomatic combinations of verbs with prepositions or postpositions (IAVs) were separated from verb-particle constructions (VPCs) such as *put off, to blow up, to do in*, in which the particle completely changes the meaning or adds a partly predictable but non-spatial meaning to the verb. Unlike VPCs, which are characteristic of Germanic languages and Hungarian, less so of Romance languages, and absent in Slavic languages, IAVs can exclusively be found in the Balto-Slavic language group.

In the training corpus, IAVs account for approximately 20% of all annotated VMWEs (see Table 1). Among the 710 examples, 154 diverse IAVs were identified. The following examples appear with a frequency of at least 20: *iti za* 'to be about' (always in the third person singular – *gre za*), *priti do* 'to occur', *ukvarjati se z* 'to work on sth', *vplivati na* 'to influence', *skrbeti za* 'to take care of', *temeljiti na* 'to be based on', *naleteti na* 'to encounter', *veljati za* 'to be considered' and *biti proti* 'to be against'. As per the guidelines, the IAV category also includes verb phrases that consist of an inherently reflexive verb (see 5.1) and a lexicalised prepositional morpheme (*nanašati se na* 'to refer to sth').

The most frequent lexicalised prepositional morpheme is *za* 'for', occurring with 34 different verbs (e.g. *gre za* 'to be about'), followed by *na* 'on', occurring with 33 different verbs (e.g. *vplivati na* 'to influence'). Frequent prepositional morphemes are also *z/s* 'with', *do* 'to' and *v* 'in'.

The lexicalised prepositional morpheme is usually positioned after the verb, which is true in 86% of the annotated examples. In the vast majority of cases, the morpheme is positioned directly after the verb or in a narrow window (+3 words). An exception is *gre za*, where an intervening element serves to reference preceding information (*gre [v tem primeru] za* 'it [in this case] is about'). In less frequent examples where the prepositional morpheme is positioned before the verb, the distance between the verb and the morpheme is significantly larger (in 20% of the cases, the distance is 3+ words).

Verbs with a lexicalised prepositional morpheme can also be identified based on common semantic features, e.g. the expression of (a) function or quality: *veljati za* [*favorita*] 'to be considered [a favorite]',[8] *imenovati [direktorja]* 'to name [smn a director]', *označiti za [laž]* 'to call [sth] out as [a lie]'; (b) (dis)agreement: *biti za/proti [globalizacijo]* 'to be for/against [globalization]'; (c) basis: *temeljiti na (dejstvu)* 'to be based on [fact]', *graditi na (zaupanju)* 'to build on [trust]'; (d) beginning or change of action/state: *pasti v [komo]* 'to fall in [a coma]', *prerasti v (ljubezen)* 'to blossom into [love]'; (e) change of quality or form: *pretvoriti v (energijo)* 'to convert into [energy]'; (f) survival: *iti skozi (proces)* 'to go through [a process]'; (g) active participation: *ukvarjati se z* 'to work on sth', *skrbeti za* 'to take care of sth'.

IAVs are characterized by the fact that the presence of the prepositional morpheme often changes the valency qualities of the verb, e.g. (a) when the original intransitive verb becomes transitive, as in the example *živeti* 'to live' : *živeti od koga/česa* 'to live off of sth'; (b) when there is a change in the case of the prepositional complement, e.g. *obrniti se na koga* 'to turn to someone (fig.) : *obrniti se h komu* 'to turn to someone (lit.)'. There are also many examples of movement verbs that as IAVs change meaning to a non-spatial judgment of state (*priti skozi* 'to go through' in the sense of 'to survive'). With verbs featuring a wide semantic range, the prepositional morpheme typically narrows down the meaning (*biti* 'to be' : *biti za* 'to be for, to support sth'). Some verbs within IAVs require an abstract object, e.g. *pasti v [depresijo, vrtinec nizkotnosti]* 'to fall

8 With IAVs, we also list typical collocates from the Gigafida Corpus of Written Slovene to ease semantic disambiguation.

into [depression, a whirlpool of insidiousness]', *dišati po* [*prevari*] 'to smell of [deceit]', *pokati od* [*veselja*] 'to be bursting of [joy]'.

Identifying inherently adpositional verbs poses a challenge both for human annotators and language technology tools as additional elements can intervene between the lexicalised morpheme and the verb. In addition, numerous verb-preposition combinations can denote a literal meaning while not exhibiting any change in the case of the object complement (*stati za* [*vrati*] 'to stand behind the door' : *stati za* [*dejanji*] 'to stand by one's actions'). They can also be polysemous (*priti do* [*spremembe*] 'to occur [change]' : *priti do* [*denarja*] 'to get [money]'). The analysis offers a starting point for the automatic identification of IAVs and provides possibilities for more detailed research, especially in terms of valency, sentence patterns and the semantic features of participants.

Verbal Idioms (VID)

The PARSEME Guidelines define verbal idioms (VID) as the combination of two lexicalised elements in which the verb is the syntactic head that requires at least one participant in the sentence pattern. The participants can take different syntactic roles, e.g. a direct or prepositional object complement (*plačati ceno* 'to pay a price', *zravnati z zemljo* 'to level with the earth'), a subject (*zgodba se ponavlja* 'lit. the story repeats itself'), an adverbial (*spati kot ubit* 'lit. to sleep like a dead person') or a subordinate clause (*vedeti, koliko je ura* 'lit. to know what time it is' in the sense 'to know what is going on'). VIDs must also keep a meaning that is independent of the meanings of their elements even with certain syntactic conversions. The Guidelines mention that the elements can appear in expected paradigms (declensions), in different tenses, in active or passive voices, with lexical variance, etc.

The definition provided by the PARSEME Guidelines differs from the one found in Slovene linguistics in that it focuses on the verb as the head and the lexicalised elements within the verb's sentence pattern. On the other hand, Slovene linguistics focuses primarily on the ability of the verb phrase as a whole to take the role of the predicate (Toporišič 1973/74; Kržišnik 1994). From this point of view, it is problematic to treat phrases that feature a verb as the fixed part, but as a whole do not always take the role of the predicate. In some cases, they can take the role of an object complement (*[ne spodobi se] voditi za nos* 'lit. [it is not proper] to lead someone by the nose' in the sense 'fooling someone is frowned upon'), a sentence (*srce se trga* [*komu*] '[someone's] heart is breaking'), or an adverbial (*hočeš nočeš* 'like it or not').

In the training corpus, 724 units were categorized as VIDs, which represents 22% of all VMWEs (see Table 1). As can be expected, VIDs occurring more than 10 times feature the verbs *biti* 'to be' and *imeti* 'to have'. Several other VIDs occur more than 5 times (*biti kos* 'to be sth's match', *priti prav* 'to come in handy', *igrati vlogo* 'to play a role', *pustiti pri miru* 'leave sth be', *priskočiti na pomoč* 'to rush to smn's aid', and *imeti opravka*

s/z 'to busy oneself with'), along with fixed discourse markers (cf. Dobrovoljc 2017): *se pravi* 'which is to say', *kdo ve* 'who knows'.

As mentioned above, the most frequent structures are combinations of the verb *biti* 'to be' and an adverb/adjective/noun. Taking into account their structural fixedness and semantic vagueness of the verb, they should be treated as separate lexicon entries: *biti všeč/res/mar/prida/prav/kos* 'to be likeable/true/to care/to be of benefit/to be right/to be smn's match'. This group includes phrases with a semantically wide verb *imeti* 'to have': *imeti prav/rad* 'to be right/to love', *ne imeti pojma/smisla* 'to have no clue/meaning'.

Another frequent structure in the training corpus is the combination of a verb and a noun or noun phrase. Among the verbs, the most frequent are *delati* 'to make' (*delati družbo/gužvo/izjeme/preglavice/razlike/sceno/škodo* 'to do/make company/a crowd/an expection/trouble/a difference/a scene/damage') and *dati* 'to give' (*dati polet/pečat* 'to give momentum/to leave a mark'). The latter structurally coincide with LVCs, but cannot be converted in the same way as LVCs to express possession (*Miha ima predavanje* 'Miha holds a lecture' ® *Mihovo predavanje* 'Miha's lecture', but not *Miha dela gužvo* 'Miha is crowding the place' ® **Mihova gužva* 'Miha's crowd'). The largest percentage in the training corpus is covered by VIDs consisting of a verb and a prepositional phrase. Again, the most frequent verb is *biti* 'to be' (*biti na dosegu roke* 'to be in reach', *biti na razpolago/voljo* 'to be at one's disposal'), followed by e.g. *priti* 'to come' (*priti na dan* 'to come to light', *priti na misel* 'to come to mind') and *dati* 'to give' (*dati na izbiro* 'to give a choice'). In terms of fixedness, some combinations of a verb and a nominal/prepositional phrase require an obligatory negation (*ne moči si kaj* 'can't help but', *ni ne duha ne sluha o (kom/čem)* 'no trace of sth', *ni para (komu)* 'someone has no equal').

The training corpus also features other structures, but with lower frequencies (*solze stopijo v oči (komu)* 'someone's eyes are watering', *časi se spreminjajo* 'times are changing'). These also include idioms (*bolje preprečiti kot zdraviti* 'lit. better to prevent than to cure') and comparisons (*igrati se [s kom/čim] kot mačka z mišjo* 'lit. to play [with smn/sth] like a cat plays with a mouse'), as well as verb-adverb combinations (*priti skupaj* 'to come together', *daleč priti* 'to come far') and combinations of a verb and a pronominal morpheme (*zagosti jo (komu)* 'to create mischief for someone').

Within their sentence patterns, VIDs open up predictable syntactic slots filled by participants with typical semantic roles. A quick overview of the annotated examples shows that certain verb forms are fixed or more frequent (e.g. third person or negated forms) and that lexical elements in a certain slot are to some extent predictable: *(svet, življenje, vse) postaviti na glavo* 'to turn [the world/life/everything] upside down').

Discussion and Conclusion

The conducted annotation task has shown that the annotation set-up (including the tool and the annotation scheme) is suitable. However, content-wise, the task is relatively complex and requires a more advanced linguistic background. The categories provided in the available guidelines are attributable and formalistically distinguishable from each other; categorization problems occur mostly when distinguishing collocations from VMWEs. The quantitative analysis shows that all categories are robust and present in authentic texts.

Based on the annotated VMWEs, we were able to identify certain pattern features on the syntactic and semantic levels. These patterns represent a good starting point for a set of rules for the automatic extraction of VMWEs and further language description. Methodologically, we made a shift in focus from a functional-syntactic perspective to the description of interconnected features on the morphosyntactic, syntactic, semantic, and lexical levels.

As expected, VMWEs are typically formed by verbs with a wide semantic range, e.g. *biti* 'to be', *dati* 'to give', *imeti* 'to have', which makes them lose their lexical qualities, but keep their morphological features, syntactic function, and position in the sentence pattern. The degree to which the meaning of the verb as an element of the MWE contributes to the meaning of the whole is often difficult to determine, one of the reasons being that numerous verb phrases structurally coincide with several categories, but denote no idiomatic meaning. In the text, they are difficult to distinguish from free phrases or collocations (frequent, semantically sensible and structurally adequate word co-occurrences).

On the other hand, the initial structural and semantic analysis has shown that (a) individual types of VMWEs form recognizable structural patterns, e.g. verb + nominal/prepositional phrase; (b) the lexicalization of elements influences the change in the participants' position and their semantic roles (*vreči se po kom* 'to take after smn' – *vreči se v kaj* 'to begin working enthusiastically' – *vreči koga ven* 'to throw smn out'); (c) that the sequence of verb elements in a VMWE is usually not fixed, but (e) there are certain tendencies in word order and (d) the number and representation of intervening elements. Furthermore, (e) certain lexical elements can be predicted based on the frequency and the elements of the co-text; (f) for better automatic identification of VMWEs, their formalized description should include information on all levels of language description.

The list of VMWEs obtained from the annotated corpus represents a set of lexicon units that can be used in machine learning for the automatic identification of VMWEs in text.

While our research did not include a systematic analysis of the sentence patterns, it should be mentioned that the training corpus includes the syntactic (formalized syntactic dependencies) and semantic (semantic role labeling) data that can be used to analyze them. This would allow us to identify more general sentence patterns for a certain VMWE type and use them in automatic extraction.

To correctly identify different MWEs, we will also create a typology of non-verbal MWEs, e.g. nominal (*žlahtna kapljica* 'fine wine'), adjectival (*vreden greha* 'worthy of sin'), or adverbial phrases (*zdaj ali nikoli* 'now or never'), as well as phrases containing particles, conjunctions and pronouns (*ja pa ja* 'as if', *s tem da* 'taking into account that') which were identified as frequent n-grams (Dobrovoljc 2017). Another challenge to tackle is the relation between the canonical and converted forms of MWEs, e.g. začarani krog 'vicious circle' – *biti ujet v začarani krog/v začaranem krogu* 'to be caught in a vicious circle' – *izviti se/rešiti se iz začaranega kroga* 'to escape from a vicious circle' – *vrteti se/znajti se v začaranem krogu* 'to spin/end up in a vicious circle' – *izstopiti iz začaranega kroga* 'to step out of a vicious circle', etc. Furthermore, it is difficult to identify MWEs with an independent, but non-metaphorical meaning, e.g. fixed expressions of the type *tehnološki park* 'technological park' and *ustavno sodišče* 'supreme court', which are closer to terminology and named entities.

Acknowledgments

The authors acknowledge the financial support from the Slovenian Research Agency: (a) research core funding No. P6-0215, *Slovene Language – Basic, Contrastive, and Applied Studies*; (b) research core funding No. P6-0411, *Language Resources and Technologies for Slovene*; and (c) project funding No. J6-8256, *New grammar of modern standard Slovene: resources and methods*. The research was conducted within the framework of the IC1207 PARSEME COST Action[9] and the IS1305 ENeL COST Action.[10]

Sources and Literature

- Arhar Holdt, Špela, and Vojko Gorjanc. 2007. "Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa." *Jezik in slovstvo* 52, No. 2 (January): 95–110.
- Atkins, Sue B. T., and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography.* New York: Oxford University Press.
- Baldwin, Timothy, and Su Nam Kim. 2010. "Multiword Expressions" In *Handbook of Natural Language Processing*, edited by Nitin Indurkhya and Fred J. Damerau, Second Edition, 267–92. Boca Raton: CRC Press.
- Breznik, Anton. 1916. *Slovenska slovnica za srednje šole*. Celovec: Družba sv. Mohorja.
- Breznik, Anton. 1934. *Slovenska slovnica za srednje šole*. 4th, enlarged edition. Celje: Družba sv. Mohorja.
- Candito, Marie, Fabienne Cap, Silvio Cordeiro, Vassiliki Foufi, Polona Gantar, Voula Giouli, Carlos Herrero, Mihaela Ionescu, Verginica Mititelu, Johanna Monti, Joakim Nivre, Mihaela Onofrei, Carla Parra Escartín, Manfred Sailer, Carlos Ramisch, Monica-Mihaela Rizea, Agata Savary, Ivelina Stonayova, Sara Stymne, Veronika Vincze. 2016. *PARSEME Shared Task 1.0 Annotation Guidelines – version 1.6b* – last updated on November 26, 2016. http://parsemefr.lif.uiv-mrs.fr/parseme-st-guidelines/1.0/.

9 *Home – PARSEME*, http://www.parseme.eu.

10 *Action IS130S – COST*, www.elexicography.eu.

- Dobrovoljc, Kaja. 2017. "Multi-word Discourse Markers and Their Corpus-driven Identification: the Case of MWDM Extraction from the Reference Corpus of Spoken Slovene." *International Journal of Corpus Linguistics* 22, No. 4 (December): 551–82.
- Dobrovoljc, Kaja, Simon Krek, and Jan Rupnik. 2012. "Skladenjski razčlenjevalnik za slovenščino." In *Zbornik Osme konference Jezikovne tehnologije*, edited by Tomaž Erjavec and Jerneja Žganec Gros, 42–47. Ljubljana: Jožef Stefan Institute.
- Gantar, Polona, Lut Colman, Carla Parra Escartín and Héctor Martínez Alonso. 2018. "Multiword Expressions: Between Lexicography and NLP." *International Journal of Lexicography*: 1–25.
- Gantar, Polona, Špela Arhar Holdt, Jaka Čibej, Taja Kuzman, and Teja Kavčič. 2018. "Glagolske večbesedne enote v učnem korpusu ssj500k 2.1." In *Proceedings of the Conference on Language Technologies & Digital Humanities*, edited by Darja Fišer and Andrej Pančur, 85–92. Ljubljana: Znanstvena založba Filozofske fakultete.
- Gantar, Polona, Simon Krek, and Taja Kuzman. 2017. "Verbal Multiword Expressions in Slovene." *Europhras 2017, Computational and Corpus-Based Phraseology: Proceedings*, edited by Ruslan Mitkov, 247–59. Cham: Springer.
- Godec Soršak, Lara. 2013. "Glagoli z oslabljenim pomenom v Slovarju slovenskega knjižnega jezika." *Slavistična revija* 61, No. 3 (March): 507–22.
- Gorjanc, Vojko, Polona Gantar, Iztok Kosem, and Simon Krek, eds. 2017. *Dictionary of Modern Slovene: Problems and Solutions*. Ljubljana: Ljubljana University Press, Faculty of Arts. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/book/15.
- Grčar, Miha, Simon Krek, and Kaja Dobrovoljc. 2012. "Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik." In *Zbornik Osme konference Jezikovne tehnologije,* edited by Tomaž Erjavec and Jerneja Žganec Gros. Ljubljana: Jožef Stefan Institute.
- Krek, Simon, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar, and Taja Kuzman. 2017. "Training Corpus Ssj500k 2.0." *Slovenian Language Resource Repository CLARIN.SI*. http://hdl.handle.net/11356/1165.
- Kržišnik, Erika. 1994. "Slovenski glagolski frazemi (ob primeru glagolov govorjenja)." PhD diss., Faculty of Arts, University of Ljubljana.
- Metelko, Franc Serafin. 1825. *Lehrgebäude der slowenischen Sprache im Königreiche Illyrien und in den benachbarten Provinzen*. Laibach: Leopold Eger.
- Ramisch, Carlos, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar et al. 2018. "Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions." In *Proceedings: LAW-MWE-CxG 2018, The 12th Linguistic Annotation Workshop (LAW XII) and the 14th Workshop on Multiword Expressions (MWE 2018)*, edited by Agata Savary, Carlos Ramisch, Jena D. Hwang, Nathan Schneider, Melanie Andresen, Sameer Pradhan, and Miriam R. L. Petruck, 222–40. Santa Fe: Association for Computational Linguistics. http://aclweb.org/anthology/W18-49.
- Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. "Multiword Expressions: a Pain in the Neck for NLP." In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics* (CICLing 2002), edited by Alexander Gelbukh, 1–15. Berlin, Heidelberg, New York: Springer.
- Schneider, Nathan, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014. "Comprehensive Annotation of Multiword Expressions in a Social Web Corpus." *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014),* edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis, 455–61. European Languages Resources Association (ELRA).
- *Slovar slovenskega knjižnega jezika*. 2nd edition. Ljubljana: SAZU and Fran Ramovš Institute of the Slovenian Language ZRC SAZU. www.fran.si.
- Toporišič, Jože. 1973/74. "K izrazju in tipologiji slovenske frazeologije." *Jezik in slovstvo* 19, No. 8 (Spring): 273–79.

- Toporišič, Jože. 1982. *Nova slovenska skladnja*. Ljubljana: Državna Založba Slovenije.
- Toporišič, Jože. 2000. *Slovenska slovnica*. Maribor: Založba Obzorja.
- Vidovič-Muha, Ada. 1998. "Pomenski preplet glagolov imeti in biti – njuna jezikovnosistemska stilistika." *Slavistična revija* 46, No. 4: 293–323.
- Žele, Andreja. 1999. "Vezljivost v slovenskem knjižnem jeziku (s poudarkom na glagolu)." PhD diss., Faculty of Arts, University of Ljubljana.
- Žele, Andreja. 2002. "Prostomorfemski glagoli kot slovarska gesla." *Jezikoslovni zapiski* 8, No. 1: 95–108.
- Žele, Andreja. 2012. *Pomensko-skladenjske lastnosti slovenskega glagola.* Linguistica et philologica 27. Ljubljana: Založba ZRC, ZRC SAZU.

Polona Gantar, Špela Arhar Holdt, Jaka Čibej, Taja Kuzman

Structural and Semantic Classification of Verbal Multi-Word Expressions in Slovene

SUMMARY

In the paper, we present an analysis of Slovene verbal multi-word expressions (VMWEs) based on the categorization made within PARSEME COST Action Shared Task 1.1 for 20 different languages. The purpose of the task was to identify VMWEs in running text based on syntactic and semantic guidelines, as well as to compile a manually annotated multi-language corpus to be made available under a Creative Commons licence. The results of the analysis will be useful in the compilation of a digital lexicon of Slovene multi-word units and will help establish a theoretical framework that takes into account the specific characteristics of Slovene while still fulfilling international criteria.

Unlike the functional-syntactic criteria advocated thus far in Slovene studies (Toporišič 1973/74; Kržišnik 1994), the classification of VMWEs within the PARSEME Shared Task 1.1 focuses on the identification of the syntactic head of the MWE. This allows MWEs to be divided into e.g. verbal, adjectival, and nominal MWEs regardless of the function they have in the sentence as a semantic and syntactic whole. The PARSEME classification consists of both universal and language-specific categories. Universal categories include verbal idioms (VID; *plačati ceno* 'to pay the price') and light verb constructions, which are further divided into full (LVC.full; *imeti mnenje* 'to have an opinion') and causal (LVC.cause; *spraviti v smeh* 'to make smn laugh'). Language-specific categories encompass inherently reflexive verbs (IRV; *zdeti se* 'to seem'), which are typical of most Slavic languages; phrasal verbs (VPC), typical of Germanic languages; and inherently adpositional verbs (IAV), also typical of most Slavic languages, including Slovene. A total of 13,511 sentences in the Slovene training corpus ssj500k 2.0 (Krek et al. 2017) were annotated with 3,364 VMWEs: 1,627 IRV

(48%), 724 VID (22%), 710 IAV (21%), 239 LVC.full (7%), and 64 LVC.cause (2%).

A linguistic analysis of the individual categories highlights numerous semantic and syntactic characteristics of the identified VMWEs that can be taken into account in the compilation of a MWE lexicon and the automatic identification of MWEs in text. Among other things, the results show the importance of the criteria used to distinguish between different types of reflexive verbs based on the role of the reflexive pronoun; they can be viewed either as independent lexical units with their own meaning (e.g. *delati se* 'to pretend') or as verbal phrases denoting e.g. mutual (*poljubljati se* 'to kiss each other'), reflexive (*umivati se* 'to wash oneself'), or passive actions (*ponavljati se* 'to be repeated'). The analysis has also shown that although the order of the components of a VMWE is usually not fixed, certain tendencies exist in terms of word order and the number of intervening elements. A semantic analysis of VMWEs has also revealed the presence of semantic groups formed by VMWEs within an individual category, as well as the properties of light verbs and verbs that typically form idiomatic units.

The study provides a good basis for further analyses of Slovene MWEs. In the training corpus, VMWE annotations can be analyzed in terms of their formalized syntactic dependency trees or the semantic roles played by the participants in the sentence.

Polona Gantar, Špela Arhar Holdt, Jaka Čibej, Taja Kuzman

STRUKTURNA IN POMENSKA KLASIFIKACIJA GLAGOLSKIH VEČBESEDNIH ENOT V SLOVENŠČINI

POVZETEK

V prispevku predstavljamo analizo glagolskih večbesednih enot (GVBE) v slovenščini na podlagi kategorizacije, kot je bila izdelana v okviru PARSEME COST Action Shared Task 1.1 za 20 različnih jezikov. Namen naloge je bil identificirati GVBE v tekočem besedilu na podlagi skladenjskih in pomenskih smernic ter izdelava ročno označenega večjezičnega korpusa, ki bo na voljo pod licenco Creative Commons. Rezultati analize bodo uporabljeni pri izdelavi digitalnega leksikona večbesednih enot za slovenščino kot tudi za utemeljitev teoretičnih izhodišč, ki upoštevajo specifike slovenščine in so hkrati usklajena z mednarodnimi merili.

Klasifikacija VMWE znotraj Parseme Shared task 1.1 za razliko od funkcijsko-skladenjskih meril, ki jih predvideva slovenistično jezikoslovje (Toporišič 1973/74; Kržišnik 1994), postavlja v izhodišče prepoznavanje skladenjskega jedra MWE, kar omogoča njihovo delitev na glagolske, pridevniške, samostalniške ipd. GVBE, neodvisno od funkcije, ki jo v stavku opravljajo kot pomenska in skladenjska celota. V izhodišču predvideva Parsemovska klasifikacija univerzalne in jezikovnospecifične

kategorije. Znotraj prvih loči glagolske idiome (VID; *plačati ceno*) in zveze z glagoli v pomensko oslabljeni rabi, ki so členjeni na prave (LVC.full; *imeti mnenje*) in vzročne (LVC.cause; *spraviti v smeh*). Znotraj druge skupine pa inherentno povratne glagole (IRV; *zdeti se*), ki so tipični za večino slovanskih jezikov, frazne glagole (VPC), značilne za germanske jezike, in glagole z leksikaliziranim predložnim morfemom (IAV), ki so tipični za slovenščino in večino slovanskih jezikov. V učnem korpusu ssj500k 2.0 (Krek et al. 2017) smo označili 13,511 stavkov, v katerih smo identificirali skupno 3,364 VMWE v naslednjih deležih: 1,627 IRV (48 %), 724 VID (22 %), 710 IAV (21 %), 239 LVC.full (7 %) in 64 LVC.cause (2 %).

Jezikoslovna analiza posameznih kategorij je pokazala številne semantične in skladenjske značilnosti identificiranih GVBE, ki jih bo mogoče upoštevati pri izdelavi leksikona VBE ter pri njihovi avtomatski identifikaciji v besedilu. Med drugim je izpostavila merila za ločevanje različnih tipov povratnih glagolov na podlagi vloge povratnega zaimka, kar omogoča njihovo obravnavanje bodisi kot samostojnih leksikalnih enot z lastnim pomenom (npr. *delati se*) bodisi kot glagolskih zvez v različnih upovedovalnih vlogah, kot so npr. vzajemnost (*poljubljati se*), povratnost (*umivati se*), pasivizacija (*ponavljati se*) ipd. Analize so tudi pokazale, da zaporedje elementov v GVBE navadno ni ustaljeno, obstajajo pa določene tendence glede besednega reda in števila vrivajočih se elementov. Analiza GVBE s semantičnega vidika je pokazala navzočnost določenih semantičnih skupin, ki jih tvorijo GVBE v posamezni kategoriji, kot tudi lastnosti glagolov v pomensko oslabljeni rabi ter glagolov, ki tipično tvorijo idiomatične enote.

Raziskava postavlja dobre osnove za nadaljnje analize VBE v slovenščini, zlasti ob upoštevanju skladenjskih oznak v obliki formaliziranih skladenjskih drevesnic v učnem korpusu, in semantičnih vlog, pripisanih udeležencem v stavčnem vzorcu.

1.01 UDC: 004.934:821.163.41

Aniko Kovač,[*] Maja Marković[**]

A Mixed-principle Rule-based Approach to the Automatic Syllabification of Serbian

IZVLEČEK

MEŠANI PRISTOP K AVTOMATSKEMU ZLOGOVANJU V SRBŠČINI NA PODLAGI NAČEL IN PRAVIL

V tem prispevku predstavljamo mešani pristop k avtomatskemu zlogovanju v srbščini na podlagi načel in pravil, ki temelji na predpisnih pravilih tradicionalne slovnice v kombinaciji z načelom zaporedja glede na zvočnost (Sonority Sequencing Principle). Proučujemo težave in omejitve obeh uveljavljenih pristopov, ki temeljita na zbirki pravil in zvočnosti; vpeljujemo algoritem, ki uporablja oba načina za doseganje natančnejše členitve besed na zloge, ki bi bila skladnejša z intuicijo rojenih govorcev; in predstavljamo statistične podatke, povezane z razporeditvijo zlogov in njihovo strukturo v srbščini.

Ključne besede: zlog, pristop na podlagi pravil, zvočnost, računalniško jezikoslovje, fonologija

ABSTRACT

In this paper we present a mixed-principle rule-based approach to the automatic syllabification of Serbian, based on prescriptive rules from traditional grammar in combination with the Sonority Sequencing Principle. We explore the problems and limitations of the existing rule set and sonority-based approaches, introduce an algorithm that utilizes both means in an attempt to produce a more accurate segmentation of words into syllables that is

[*] **Department of Language Science and Technology, Saarland University Campus A2 2, 66123 Saarbrücken, Germany, anikok@coli.uni-saarland.de**

[**] **Department of English Language and Literature, Faculty of Philosophy, University of Novi Sad, Dr Zorana Đinđića 2, 21000 Novi Sad, Serbia, majamarkovic@ff.uns.ac.rs**

better aligned with the intuition of the native speakers, and present the statistical data related to the distribution of syllables and their structure in Serbian.

Keywords: syllable, rule-based approach, sonority, computational linguistics, phonology

Introduction

Syllables have been considered — although not unequivocally (cf. Koehler 1966) — to be one of the basic units in phonology constituting the minimal units of pronunciation, and to play a role in prosody, phonotactics, and phonological processing (Ladefoged and Johnson 2014). The role of the segmentation of words into syllables and their distributional properties began to see an increase in importance in speech technologies in the 1990s (Iacoponi and Savy 2011), most notably in the areas of speech recognition (SR) and text-to-speech synthesis (TTS).

Syllable segmentation today plays a role in speech technologies on the segmental level — conditioning the length of segmental units such as consonants and vowels — as well as on the prosodic level — governing rhythmical alternations (Bigi and Petrone 2014). Syllable segmentation is also a key component in hyphenation (e.g. Kaplar et al. 2018), although it should be noted that, at least in Serbian, hyphenation is governed by a partially diverging set of rules from those governing syllabification[1]. Syllable distribution data is also of crucial importance for psycholinguistic experiments, as syllable frequency has been shown to play a role in the processing of words (e.g. Barber et al. 2004; Cholin et al. 2006; Cholin and Levelt 2009). Developing an automatic system of syllabification allows for the segmentation of large-scale language corpora needed for the development of automatic systems or the extraction of relevant data related to frequency syllable distributions, which would otherwise require a large number of trained annotators and would be a resource and cost heavy undertaking.

The two generally distinguishable approaches to automatic syllabification are rule-based versus data-driven approaches (Marchand et al. 2009). While data-driven approaches have taken over many aspects of natural language processing, and there are a number of data-driven models of syllable segmentation using artificial neural networks (e.g. Daelemans and van den Bosch 1992; Hunt 1993; Stoianov et al. 1997; Landsiedel et al. 2011), the unavailability of segmented data for Serbian makes rule-based approaches the only viable option for automatic syllabification in Serbian.

To the best of our knowledge, there is a single publicly available attempt at developing a rule-based syllabifier for Serbian by Kaplar et al. (2018). In this paper we lay out a number of problems and limitations with the ruleset used in their syllabification system and why relying on the existing set of prescriptive rule descriptions from traditional grammar is insufficient to capture and describe a syllabification system that

1 For example, hyphenation rules ban the segmentation after a syllable consisting of a single vowel at word onset, while this segmentation is allowed and expected according to the rules of syllabification.

is aligned with the intuition of native speakers of Serbian. A relatable attempt at automatic syllabification was developed by Meštrović et al. (2015) for Croatian, the key difference between their work and ours being in the principle behind the syllabification algorithm which in their case relied solely on the onset maximization principle — limiting possible syllable onsets to valid onsets at the beginning of words. Taking into account Morelli's (1999) limitations on possible syllable onsets in Serbo-Croatian, the onset maximization principle employed by Meštrović et al. could be considered a comparatively liberal system. In order to attempt to constrain our syllabifer, we are decided on a different approach that will not rely on onset maximization, but rather a combination of a number of alternative principles.

In this paper we present a mixed-principle rule-based approach to the syllabification of Serbian. Our starting set of rules is based on the *Gramatika srpskoga jezika* by Stanojčić and Popović (2005), a prescriptive textbook for Serbian grammar that presents a set of rule descriptions for the segmentation of words into syllables. In a previous version of our syllabification algorithm (Kovač and Marković 2018), we made a number of changes to the rule descriptions of Stanojčić and Popović (2005) as the formulation of some of the descriptions proved to be redundant, some were example-based and not specific enough for a formal implementation, and we also expanded them with three added modifications related to the treatment of nasals and the alveolar sonorant /*r*/ based on Kašić (2014) and the treatment of alveolar sonorants /*l*/ and /*n*/ based on Zec (2000). In this paper we extend our previous algorithm to include a module for validating the structure of syllables in terms of their compliance with the Sonority Sequencing Principle (SSP), thus further fine-tuning the accuracy of our segmentation, and resolving a number of problems noted in our earlier implementation.

The goal of the paper is threefold: i) to improve our system for automatic rule-based syllabification for Serbian based on the formalization of existing rule descriptions by the addition of the sonority sequencing validation module, ii) to provide an analysis of the outcomes of the automatic syllabification process in order to address possible theoretical considerations and serve as a basis for the development of future syllabifiers, and iii) to present statistical data related to the distribution of syllables and their structure in Serbian.

Prescriptive Rule Descriptions

Our starting set of rules was based on the formalization of the rule descriptions governing the segmentation of words into syllables from the *Gramatika srpskoga jezika* by Stanojčić and Popović (2005). Being a prescriptive textbook on Serbian grammar used at a high school level by all student profiles, we expected these rules to constitute the common knowledge base shared by the majority of native speakers.

Regarding syllable boundaries, Stanojčić and Popović (2005, 37) establish the following general rule (1).

(1) *In words made up of multiple phonemes, consonants, sonorants and vowels, the syllable boundary comes after the vowel and before the consonant (e.g.* či-ta-ti *[*to read*]).*

In addition to this general rule, they list the following rules — (2), (3), (4), (5) and (6) — that further specify medial syllable boundaries depending on consonant manner of articulation.

(2) *Medially, in a consonant cluster which has an affricate or fricative sound in its initial position, the syllable boundary will be before that consonant cluster (e.g.* po-šta *[*post*]*, ma-čka *[*cat*]).*
(3) *The syllable boundary will be before a consonant cluster if, in a consonant cluster found medially in a word, the second position in the cluster is occupied by one of the sonorants /*v*/, /*j*/, /*r*/, /*l*/ or /*ʎ*/ preceded by any other consonant besides a sonorant (e.g.* sve-tlost *[*light*]).*
(4) *If a consonant cluster consists of two sonorants, the syllable boundary will be between them so that one sonorant belongs to the preceding, and one sonorant belongs to the following syllable (e.g.* lom-ljen *[*broken*]).*
(5) *If a consonant cluster consists of a plosive in its initial position and some other consonant except the sonorants /*j*/, /*v*/, /*l*/, /*ʎ*/ and /*r*/, the syllable boundary will be between the consonants (e.g.* lep-tir *[*butterfly*]).*
(6) *If in a cluster of two sonorants, the second position is occupied by the sonorant /*j*/ from je corresponding to the ijekavica dialect to /*e*/ in the ekavica dialect, the syllable boundary will be before that group (e.g.* čo-vjek *[*man*]).*

Stanojčić and Popović (2005, 32) also introduce the rule descriptions (7) and (8) to define when the sonorants /*r*/, /*l*/, and /*n*/ constitute syllable nuclei.

(7) *The sonorant /*r*/ can be a syllable carrier in standard Serbian when:*
   a. it is found medially between two consonants (e.g. tr-ča-ti *[*to run*]),*
   b. it is found initially before a consonant (e.g. r-va-ti se *[*to wrestle*]),*
   c. it is found after a vowel in compounds (e.g. za-r-đa-ti *[*to rust*]),*
   *d. before /*o*/ that is realized as an /*l*/ in other members of the paradigm (e.g.* o-tr-o *(m.) from* o-tr-la *(f.) [*wiped*]).*
(8) *The other two alveolar sonorants, /*l*/ and /*n*/ can be syllable carriers in dialectal toponyms (e.g.* Stlp, Vlča glava, Žlne*) or foreign toponyms (e.g.* Vltava, Plzen*) but also in other personal names (e.g. English* Idn *or Arabic* Ibn-Saud*), and in the word* bicikl *[*bicycle*].*

Revising the Existing Rule Set

The development of our syllabification algorithm has been an iterative process testing the existing rule set and making changes as needed. While other authors (e.g. Kaplar et al. 2018) used the rule descriptions of Stanojčić and Popović (2005) directly

to implement a software architecture for syllabification in Serbian, we have found a number of problems with this approach.

The definition of the rule description under (1) causes the initial member of a consonant cluster in the rule descriptions under (2)–(6) to be understood as the first consonant following a vowel. However, given that the sonorants /*r*/, /*l*/, and /*n*/ can also constitute syllable nuclei in Serbian in certain positions, as presented under rule descriptions (7) and (8), a more precise definition would be that the initial member of a consonant cluster is the first consonant following an element that constitutes a syllable nucleus. The general rule under (1) should be then revised as follows.

(1*) *In words made up of multiple phonemes, consonants, sonorants and vowels, the syllable boundary comes after the vowel or sonorants /*r*/, /*l*/, and /n/ in syllable bearing positions and before the consonant (e.g.* či-ta-ti *[*to read*],* tr-ča-ti *[*to run*]).*

In addition to our expansion of the general rule presented under (1) to include the syllable bearing sonorants, while formalizing the rule descriptions via finite-state automata, rules (2) and (3) proved to be redundant as they produced identical outcomes to the general rule under (1*). Because of this, these rules were disregarded in our syllabification algorithm.

During our early testing of the verbatim implementation of the rule descriptions, we also noticed that the existing rule descriptions treated a consonant cluster consisting of a nasal in initial position followed by a consonant that is not one of the sonorants /*j*/, /*v*/, /*l*/, /*ʎ*/, and /*r*/ as a part of the following syllable onset, producing outcomes such as: *gu-ngula* [*commotion*], *mo-mci* [*guys*], *ka-ncelarije* [*offices*], *su-nce* [*sun*], etc. Contrary to Stanojčić and Popović (2005), authors such as Kašić (2014) argue that nasals should be treated analogously to plosives during syllabification because there is a complete occlusion in the oral cavity during their production. If this principle were to be employed, rule (5) should be revised as follows.

(5*) *If a consonant cluster consists of a plosive or nasal in its initial position and some other consonant except the sonorants /*j*/, /*v*/, /*l*/, /*ʎ*/,* and */r/, the syllable boundary will be between the consonants.*

Following rule (5*), the examples above would then be segmented as: *gun-gula* [*commotion*], *mom-ci* [*guys*], *kan-celarije* [*offices*], *sun-ce* [*sun*], etc. Even though in the earlier implementation of our syllabifier (Kovač and Marković 2018) we did not want to employ the Sonority Sequencing Principle (SSP), we opted for the treatment of nasals by Kašić (2014) in our implementation, which respected the limitations put forward by the Sonority Hierarchy, and was more in line with native speaker intuition.

The Sonority Hierarchy

Sonority Theory accounts for the organization of segments into well-formed sequences, both within the syllable and across syllabic boundaries. This organization is driven by principles of sonority, a property that is used as the basis of ranking all sounds along a scale, from less sonorous to more sonorous ones. Although there is a general consensus that segments are ranked by their inherent sonority, the notion of sonority itself is not unambiguously described in the phonetic and phonological literature. Among the phonetic approaches, Ladefoged (1982) defines sonority as the perceptual salience or loudness of a sound, and Bloch and Trager (1942; according to Goldsmith 1995) define it as the amount of airflow in the resonance chamber. For others, sonority is dependent on multiple phonetic parameters (Ohala and Kawasaki 1984; Ohala 1990; Butt 1992). In the phonological literature, sonority is generally defined as a multi-valued feature (Foley 1972; Hankamer and Aissen 1974; Selkirk 1984), although there are also authors who argue that it is derivable from the more basic binary features of phonological theory (Clements 1990). Other questions that are often addressed are whether sonority scales are universal or language-specific, allowing freedom to languages in assigning sonority values, and how fine-grained distinctions sonority scales should capture. For example, Clements' universal sonority scale includes only four major classes of consonants (Clements 1990), ranked from least sonorous to most sonorous, as in (i):

(i) O < N < L < G
(O = obstruents, N = nasals, L = liquids, G = glides)

Selkirk (1984, 112) proposes a much more detailed scale, which divides all sounds into 11 groups, assuming more subtle differences in sonority values. Selkirk also states that the sonority indices may not be as important in themselves as the sonority relations that they express. Selkirk's scale of sonority in consonants is given in (ii):

(ii) p, t, k < b, d, g < f, θ < v, z, ð < s < m, n < l < r

Sonority scales serve as the basis of constructing segment sequences within syllables. The universal cross-linguistic generalization is that in the sequence of segments, the one ranking highest on the sonority scale constitutes the peak of the syllable, i.e. it is the syllabic nucleus. As for the other segments around the nucleus, they are organized so that the more sonorous ones are closer to the nucleus, and less sonorous ones are more distant. This generalization is referred to as Sonority Sequencing Principle (SSP). Thus a syllable with an ascending sonority slope in the onset and a descending slope in the coda, such as, for example *blunt*, is a well-formed syllable, whereas **lbutn* is prohibited, due to the violation of the SSP. Adopting thee SSP often solves the problems of syllabic consonants, since they generally occur in environments where they constitute a sonority peak, as in the Serbian word *pr-vi*.

The Need for Sonority

Apart from the segmentation of nasals analogously to plosives following Kašić (2014) that relied on principles of the SSP, in our initial attempt at the formalization of the rule description under (8) of Stanojčić and Popović (2005) we had to rely on sonority to define the criteria for when the alveolar sonorants /*l*/ and /*n*/ act as syllable nuclei.

As Stanojčić and Popović gave no formal criteria defining the contexts of syllable bearing /*l*/ and /*n*/, our initial attempt to draw on generalizations based on their examples for syllable carrying /*l*/ (*Stlp, Vlča glava, Žlne, Vlava, Plzen*) and /*n*/ (*Idn, Ibn-Saud*). In analogy to the rules descriptions under (7a) and (7b) and our added rule (7c*) defining the contexts in which the alveolar phoneme /*r*/ can act as a syllable nucleus, we implemented rule (8*) to define the conditions under which the phonemes /*l*/ and /*n*/ can act as syllable bearing nuclei.

(8*) *The other two alveolar sonorants, /*l*/ and /*n*/, can be syllable carriers if they are found:*
a) medially between two consonants,
b) initially before a consonant, or
c) finally after a consonant.

However, the formulation under (8*) allowed for outcomes such as: *Be-rn, Ka-rl, erla-jn, Kla-jn, kasa-rn-skim, Linko-ln, Va-jl-om*, etc. in which the phonemes /*l*/ and /*n*/ identified as syllable nuclei have a lower sonority level than the consonants in their onset or coda. Because the phonemes /*r*/ and /*j*/ are more sonorous than the phonemes /*l*/ and /*n*/, and the lateral phoneme /*l*/ is more sonorous than the nasal phoneme /*n*/, native speakers do not perceive the elements of lower sonority as syllable nuclei in these contexts. Zec (2000) states that alveolar sonorants can be syllable bearing elements in Serbian only in contexts in which there is no segment of a higher level of sonority in their immediate vicinity. Because of this, we needed to further specify rule (8*) to take sonority constraints into consideration as follows.

(8**) *The other two alveolar sonorants, /*l*/ and /*n*/, can be syllable carriers if they are found:*
a) medially between two consonants of lower sonority,
b) initially before a consonant of lower sonority, or
c) finally after a consonant of lower sonority.

It turns out that this principle can also account for the behavior of the syllable bearing /*r*/ in Serbian. In fact, it does not only provide a general account for consonantal syllabic nuclei in Serbian that subsumes the rules under (7) and (8**) it also accounts for our extension of rule (7) that keeps the the consonant cluster /*rje*/ of

the ijekavica dialect unsegmented in initial position[2]. Because the phoneme /j/ has a higher level of sonority than /r/, the phoneme /r/ should not be treated as a syllable nucleus initially in words such as *rjeka* [*river*].

In our previous implementation of the syllabifier (Kovač and Marković 2018), we attempted to limit our reliance on the Sonority Sequencing Principle to the cases above. However, during the evaluation of our algorithm, we encountered a number of syllable structures that were unexpected due to their absence from the onset maximization approach to syllabification developed for Croatian by Meštrović et al. (2015). Namely, we encountered the syllable structure CCCCCVC in *mo-na-rhstvom* [*with the monarchy*], the structure CCCCCV in the words *se-rbska* [*Serbian*], *ca-rstva* [*kingdoms*], and *sta-ra-te-ljstva* [*custody*], and the structure CCCCVC in *se-rbskom* [*Serbian*], *de-jstvom* [*with effect*], *vo-đstvom* [*leadership*], *spo-rtskim* [*sport*], and *a-lpskog* [*alpine*].

The way we attempted to remedy this issue was to limit the syllable onset length three-syllable clusters, which is the maximum length of non-syllabic consonant clusters word initially in Serbian (Kašić 2014). While this constraint, in combination with rules (5) and (6), resolved the issues in the examples we encountered — with this limitation, they are segmented as *mo-narh-stvom* [*with the monarchy*], *serb-ska* [*Serbian*] (three-syllable onset limitation + rule (5)), *car-stva* [*kingdoms*], *sta-ra-telj-stva* [*custody*], *serb-skom* [*Serbian*], *dej-stvom* [*with effect*], *vođ-stvom* [*leadership*], *sport-skim* [*sport*], *alp-skog* [*alpine*] — some medial clusters with a syllabic consonant still remained a problem. For example, in the word *najstrpljiviji* [*most patient*], which contains a syllabic /r/, the syllable boundary that would be placed between /na/ and /jstr/ — *na-jstr-pljiviji* — which does not coincide with native speaker intuition. The Sonority Sequencing Principle seems like a perfect solution for this cases, as it would require the structure of a syllable to follow a sonority scale, with the syllable nucleus being the most sonorous element, while sonority would gradually decrease towards the periphery of the syllable (Zec 2000). With this added sonority requirement, the phoneme /j/, being more sonorous than /s/ and /t/, would have to constitute a part of the previous syllable where it would be of a lower sonority when compared to its neighbouring syllable bearing vowel, and the syllable boundary would be *naj-str-pljiviji* which is in line with native speaker intuition.

As a final check following rules (1)–(8**), we add rule (9) that has the ability to shift the syllable boundary in order to avoid a violation of the sonority hierarchy.

(9) *If the syllable structure resulting from rules (1)–(8**) does not conform to the Sonority Sequencing Principle, move the boundary so that the phoneme violating the sonority sequence is shifted into the neighboring syllable.*

2 It should be noted that while sonority sequencing accounts for the non-syllabic treatment of /r/ before /je/ in initial position, our rule extension is still needed as it has a more general scope than the sonority rule and accounts for segmentation in medial positions as well (e.g. in words such as *isko-rje-nilo* [*eradicated*]).

An Adapted Sonority Hierarchy

In our sonority sequencing module, we relied on a combination of Selkirk's (1984) sonority scale, the sonority apertures for Serbian described by Subotić et al. (2012), and some notes on sonority sequencing in Serbian from Zec (2000). Our sonority scale is shown under (iii).

(iii) p, t, k < b, d, g < ts, tʃ, tɕ < f, ʃ, h < v, z, ʒ < s < m, n, ɲ < l, ʎ < j, r < a, e, i o, u

The highest sonority group in our implementation was made up by the vowels of Serbian. As vowels constitute syllable nuclei and there can only be a single vowel per syllable, we did not need to make a distinction between three sonority apertures of vowels (i, u < e, o < a) as it is the case in the hierarchy of Subotić et al. (2012). Following Selkirk (1984), we divided sonorants into three sonority classes, and following Zec (2000), we treated liquids as more sonorous than nasals, and, within liquids, the phoneme /*r*/ as more sonorous than laterals. For the needs of our implementation, we treated the phoneme /*r*/ and glide /*j*/ as a single sonority group, although from a theoretical standpoint /*j*/ would be considered as more sonorous out of the two given its semi-vowel nature. We opted for treating /s/ as an element of higher sonority than voiced fricative despite its voiceless nature following Selkirk (1984), and expanded Selkirk's hierarchy with the addition of affricates between voiceless fricatives and voiced plosives as a parallel to the aperture order presented by Subotić et al. (2012).

It is important to note that there are sequences which clearly do not conform with the SSP in a number of languages, and which may undermine the relevance and power of the sonority hierarchy. A very common pattern, found across a number of unrelated languages, is the possibility of an /*s*/ + plosive sequence in the syllable onset, which would be in clear violation if we were to adopt the sonority scale outlined above. In Serbian, there is a known ambiguity in syllable segmentation in the case of continuant fricative phonemes. For example, the word *postaviti* [*to set*] can be syllabified as both *po-sta-vi-ti* and *pos-ta-vi-ti* (Gvozdanović 2011). We therefore adopt the view put forward in Morelli (1999), who argues that fricatives and plosives may be treated as a single class with respect to sonority in these cases — since splitting them into separate classes would make wrong typological predictions — and add an exception to our sonority sequencing module that leaves fricative + plosive sequences as a viable sequence in the syllable onset.

Our Algorithm[3]

Our mixed-principle syllabification algorithms consists of the following steps:

3 Our implementation of the algorithm can be found at https://github.com/versi-regular/rule-based_syllabifier_sr, licensed under the GNU General Public License v3.0. It was developed using Python 3.x and processes 10380 tokens/s on average estimated on a 4,681,713 token corpus processed on an Intel® Core™ i5-3210M CPU @ 2.50GHz with 8.00 GBs of DDR3L-1600 SODIMM, including pre-processing, clean-up, and transliteration.

I. Identify vowels in the word and mark their positions as positions capable of constituting syllable nuclei (based on (1)).
II. If a word contains the letters *l*, *n* or the letter *r* not followed by the sequence *je* in the center of a consonant cluster consisting of elements of lower sonority or at the beginning or a word followed by a consonant of lower sonority, or the letters *l* or *n* at the end of a word preceded by a consonant of lower sonority, treat those positions in the word as capable of constituting syllable nuclei (based on (1*), (7), and (8**)).
III. For each position identified as capable of constituting a syllable nucleus:
    A. If it is followed by a sequence of two sonorants, mark the syllable boundary between the two sonorants (based on (4)), except if the second sonorant is *j* and it is followed by *e*. If the second sonorant is *j* followed by *e*, mark the syllable boundary before the sonorant cluster (based on (6)).
    B. If it is followed by a sequence of a plosive or nasal and a plosive, fricative, affricate or nasal, mark the syllable boundary between the two consonants (based on (5*)).
    C. In all other cases mark the syllable boundary after the syllable nucleus (based on (1*)).
IV. Run a recursive sonority check (based on (9)):
    A. If the word consists of more than one syllable, convert the syllable structures identified by the previous steps into sonority group values.
    B. For each syllable, check if there is a violation of the SSP at the edges of the syllable ignoring the check at the onset on the word-initial syllable and the check in the coda of the word-final syllable.
    C. If a violation found is a sequence of a fricative followed by a plosive in the onset, ignore the violation.
    D. If there is a violation, remove the letter from the edge of the syllable, and add it onto the neighboring syllable.
    E. Repeat until no violation is found.

Syllable Distribution Data

In this section, we present the statistical distribution data of syllables in Serbian based on our updated syllabification process applied to the Serbian Lemmatized and PoS Annotated Corpus *SrpLemKor* (Popović 2010; Utvić 2011). We chose *SrpLemKor* for our analysis, because its annotation allowed us to filter out numbers, Roman numerals, abbreviations and non-Serbian words or suffixes in compounds (at least to some extent) and thus reduce noise in the data.

The following results show the syllable distribution statistics based on 3,648,543 non-unique word-forms (word tokens) from *SrpLemKor*. From a total of 4,681,713 entities (punctuation and word tokens) in our version of the corpus, 113,679 (2.43%)

entities of texts #260, #4505 and #4517 were excluded because the files contained faulty encoding. Based on corpus tags, we excluded 919,161 (19.63%) entities tagged PUNCT (punctuation), SENT (sentence separator full-stops), RN (Roman numerals), NUM @card@ (Arabic numerals), ABB (abbreviations) and ? (non-Serbian words and other uncategorized entries). An additional 815 (0.02%) entities that contained the characters w, q and x were removed in an attempt to further reduce noise stemming from foreign words, as not all foreign words were tagged as such in the corpus. In the process of syllabification, an additional 12,877 (0.28%) entities were removed as they were solely made up of consonant clusters with no available syllable nucleus candidate.

Syllable Type Distributions in Serbian

In the 3,648,543 word-forms from *SrpLemKor*, a total of 8,196,771 syllables were identified. Table 1 presents the syllable type distribution based on our mixed-principle syllabification algorithm.

Table 1: Syllable structure distribution of syllables in the *SrpLemKor* corpus

Syllable structure	No.of instances	Percent
CV	5030622	61.37321636
CCV	938275	11.44688561
CVC	913603	11.14588903
V	852854	10.40475573
CCVC	218126	2.661121068
VC	141980	1.7321455
CCCV	56168	0.685245446
CVCC	20339	0.248134296
CCCVC	14362	0.175215338
CCVCC	6274	0.076542336
VCC	2234	0.027254635
CCCCV	780	0.009515942
CVCCC	731	0.008918146
CCCVCC	170	0.002073987
CCCCVC	84	0.001024794
VCCC	67	0.000817395
CCCCVC	36	0.000439197
Other	66	0.000805195
Total	8196771	100

These results show the distribution of syllables in a somewhat noisy data. We found there are still foreign words annotated as non-foreign in the corpus constituting some of the less-frequent syllable structures listed as "Other" in Table 1. For example, an instance of the syllable structure VCCCCC was found to correspond to the segmentation of the German word *Pe-itscht* [*lashes*], the syllable structure CCCCVCCC was identified in the German word *Fle-i-schmarkt* [*meat market*], and the structure CCCCCVC was found in the German word *Gle-i-chschal-tung* [*co-ordination*]. The structure CCCCCCVC was found in the German word *Na-chtschat-ten* [*nightshade*] and in the toponym *CRYSLER*. The syllable structure CCVCCCC was found in the source transcription of the last name *Pe-tritsch* and in the English word *knights*. The syllable structure CCCVCCC was identified to be a part of the German words *Wolfsmilch* [*spurge*] and *E-in-ge-schickt* [*sent in*] and to correspond to the English word *string*. The syllable structure CCCCCCV was identified in the German words *We-ihna-chtsbra-e-u-che* [*Christmas trees*], *Stor-chschna-bel* [Crane's bill], while the structure CCCCCV was found in the words *Re-chtsge-schi-chte* [*history of law*] and *Um-gangsspra-che* [*vernacular*], as well as in the sequences *šttske* and *su-žnjstva*. The syllable structure CCCCVCC was found in the German word *Ze-it-schrift* [*magazine*], and in multiple occurrences of the source spelling of the last names *Schmidt* and *Rot-hchild*. The structure VCCCC was found in the German words *Deutsch* [*German*], *Ernst* [*seriousness*], in the sequence *der-demnaechst* [*soon*], and in the strings *ikvbv* and *EHCmc*. As can be seen from the examples above, besides foreign origin words, noise in the data can also be found in typos and strings we did not manage to identify. Another example of such string was *ngBpJKTnQ* identified as the structure VCCCCCCCC. Most structures identified as CVCCCC were the result of typos, e.g. serbsk, kra-levstv, pod-danstv, carstv, slav-jansk, ju-go-slo-venskg, cr-no-gorskg, but also foreign origin names, e.g. *Hirsch, Herbst, Lokotsch*, and *Worlds* in additions to strings such as *majnds and Gorrrr*. In addition to these, one occurrence of the syllable structure CVCCCCCCCC that stood for the onomatopoeic vulgarism *mršššššššš* [*go away*].

We also found 2 syllable structures that differed from the structures found by Meštrović et al. (2005) for Croatian. The structure CCCCVC was identified in the words *vo-đstvom* [*with leadership*], *za-ko-no-da-vstvom* [*with legislature*], *mo-nar-hstvom* [*with monkhood*], *lu-ka-vstvom* [*with slyness*], *be-zzglob-na* [*without wrists*], and in the paradigm members of the word *po-sthlad-no-ra-to-vski* [*post-cold-war*]. It also occurred in the Russian word *Zdra-vstvuj* [hello], in the German-origin word *Ha-up-tstrum-fi-rer* [*mid-level commander*], in the German *Ra-u-schmit-tel* [*intoxicant*] and *Li-e-be-spflan-ze* [*love plant*] and in the misspelled Serbian words *pri-ja-tljskih* [*friendly*] and *kvdrat* [*square*]. The structure CCCCV was found in the words *bi-vstvu* [*existence*], *va-zdu-ho-plo-vstvo* [*aviation*], *kra-lje-vstva* [*kingdoms*], *zdra-vstve-noj* [*health*], *vo-đstvo* [*leadership*], *ču-vstva* [*feeling*], *pre-i-mu-ćstva* [*advantages*], and *mo-gu-ćstvu* [*possibility*]. It also occurred in German words such as *Pfin-gstro-se* [*peony*], *Ke-u-schhe-it* [*chastity*], *Schne-e-glo-ec-kchen* [*snowdrop*], *Schne-e-ro-se* [*Chrismas rose*], *Ge-i-sskle-e* [*cystus*], *Vol-ksbra-uch* [*popular custom*], *Vol-ksgla-u-ben* [*popular belief*],

Schri-ften [*regulations*], *Schlu-e-ssel-blu-me* [*cowslip*], and more. We discuss the implications of these for our syllabification algorithm in the Discussion section below.

Syllable Type Positional Distributions in Serbian

We also examined the syllable type frequencies with respect to their position in a word. Four positional frequencies are presented in Table 2: syllable type frequencies in monosyllabic words, and syllables type frequencies in the initial position, in medial positions, and in the final position of polysyllabic words.

Table 2: Syllable structure distribution of syllables in the *SrpLemKor* corpus categorized by position

Syllable structure	Monosyllabic words		Polysyllabic words					
	MONO		INITIAL		MEDIAL		FINAL	
	No.of instances	Percent	No.of instances	Percent	No.of instances	Percent	No.of instances	Percent
CV	612214	50.382	1356771	56.064	1476732	68.956	1584905	65.49
CCV	62244	5.122	372181	15.379	305247	14.254	198603	8.21
CVC	129337	10.644	178859	7.391	211979	9.898	393428	16.26
V	301295	24.795	369133	15.253	61241	2.860	121185	5.01
CCVC	35428	2.916	50383	2.082	53397	2.493	78918	3.26
VC	64038	5.270	67539	2.791	7123	0.333	3280	0.14
CCCV	174	0.014	19754	0.816	20260	0.946	15980	0.66
CVCC	5368	0.442	1052	0.043	695	0.032	13224	0.55
CCCVC	1490	0.123	3976	0.164	4427	0.207	4469	0.18
CCVCC	1635	0.135	206	0.009	17	0.001	4416	0.18
VCC	1125	0.093	162	0.007	18	0.001	929	0.04
CCCCV	14	0.001	21	0.001	381	0.018	364	0.02
CVCCC	579	0.048	3	0.000	1	0.000	148	0.01
CCCVCC	105	0.009	0	0.000	0	0.000	65	0.00
CCCCVC	1	0.000	0	0.000	25	0.001	58	0.00
VCCC	45	0.004	0	0.000	0	0.000	22	0.00
CCCCVC	11	0.001	0	0.000	0	0.000	25	0.00
Other	38	0.003	0	0.000	7	0.000	21	0.00

Based on *SrpLemKor*, the most frequent monosyllabic syllable structures in Serbian are CV (50%), V (25%) and CVC (11%). The most frequent syllable structures in the initial position of polysyllabic words are CV (56%), CCV (15%) and V (15%). In medial positions in polysyllabic words, the most frequent syllable structures

are CV (69%), CCV (14%) and CVC (10%). The most frequent syllable structures in the final position of polysyllabic words are CV (65%), CVC (16%) and CCV (8%). It is interesting to note the asymmetry that the syllable structures CCCVCC, VCCC, and CCCCVC occurred only in monosyllabic words and in the final position of polysyllabic words, while the syllable structure CCCCVC occurred in all positions except the initial position in polysyllabic words.

Syllable Nuclei Statistics in Serbian

The distribution of different syllable nuclei in Serbian based on the *SrpLemKor* corpus is presented in Table 3.

Table 3: Syllable nuclei statistics and positional frequencies of syllables in the *SrpLemKor* corpus

Nucleus	TOTAL		Monosyllabic words		Polysyllabic words					
			MONO		INITIAL		MEDIAL		FINAL	
	No.of instances	Percent	No.of instances	Percent	No.of instances	Percent	No.of instances	Percent	No.of instances	Percent
a	2177498	26.566	330629	27.209	604764	24.990	585787	27.353	656318	27.120
e	1646579	20.088	304442	25.054	447662	18.498	394573	18.425	499902	20.657
i	1730439	21.111	230637	18.980	394735	16.311	600823	28.056	504244	20.836
l	939	0.011	326	0.027	32	0.001	77	0.004	504	0.021
n	1261	0.015	409	0.034	544	0.022	33	0.002	275	0.011
o	1753091	21.388	168126	13.836	671752	27.758	385687	18.010	527526	21.798
r	88021	1.074	1898	0.156	66250	2.738	19560	0.913	313	0.013
u	798943	9.747	178674	14.704	234301	9.682	155010	7.238	230958	9.544

Based on the positional nucleus distribution data, it can be seen that overall /a/ and /o/ constitute the most frequent nuclei in Serbian. However, there is some positional variation. While the most frequent nuclei in final, medial, and initial position of polysyllabic words are also /a/ and /o/, in monosyllabic words, the most frequent nuclei are /a/ and /e/.

Discussion

While our mixed-principle rule-based syllabification algorithm is suitable for the segmentation of words into syllables following the ruleset we devised based by the combination of prescriptive rule descriptions and the employment of the Sonority

Sequencing Principle, there are still some practical and theoretical considerations to be addressed.

While reporting on the syllable distribution data, we mentioned that the 3,648,543 word-forms extracted from *SrpLemKor* used for the calculation of statistical data related to the distribution of syllables and their structure in Serbian still contained some noise such as foreign words, typos, and possibly random character strings. Based on 500 random samples taken from the syllable output data checked by a human evaluator, the estimate of the amount of such noise in the data is <2%. Given the nature of corpus-based data, this noise should not significantly impact the reliability of the distributional information.

From a theoretical standpoint, in formulating our algorithm, we disregarded the three-syllable consonant cluster limitation put forward by Kašić (2014) in favor of exploring the limitations of the sonority module. The occurrence of the two syllable types CCCCVC and CCCCV, which were not present in the onset-maximization-based syllabification algorithm for Croatian (Meštrović et al. 2015), shows that in a limited number of instances this constraint is needed to exclude syllable clusters that are in accordance with the SSP and prescriptive rule descriptions, but seem contrary to native speaker intuition about syllable boundaries. In addition to this, there is the ambiguity in syllable segmentation in the case of continuant fricative phonemes (Gvozdanović 2011) with the continuant constituting either the first place in the onset of the syllable or the last place in the coda of the previous syllable, e.g. the possibility to syllabify *postaviti* [*to set*] as *po-sta-vi-ti* and *pos-ta-vi-ti*, would require a larger-scale study examining the intuition of native speakers on syllabification to make an assumption about contemporary tendencies in the segmentation in these contexts.

In order to verify the syllabic status of different clusters, it would be interesting to conduct a series of monitoring studies modeled after Mehler et al. (1981), who have shown that reaction times to a word are faster if the word is primed by a sequence corresponding to a syllable in the word when compared to priming with a string that does not constitute a syllable. Bradley et al. (1993) argue that these effects produce mixed results in some languages which contain a large number of ambisyllabic segments, so these studies may also reveal whether and to what extent syllables play a role in prelexical processing in Serbian.

Conclusion

In this paper we presented a mixed-principle rule-based syllabifier modelled after the rule descriptions found in Stanojčić and Popović (2005), extended by rule specifications from Kašić (2014) and Zec (2000), and complemented by a sonority sequencing module based on Selkirk (1984), Subotić et al. (2012), and Zec (2000).

An implementation of the existing prescriptive rules for the segmentation of words into syllables allowed us to gain an insight into the problem areas of the rule

descriptions, and propose a number of revisions and amendments to the existing rules. The sonority sequencing module revealed the need for an additional onset-length limitation constraint, and pointed out the limitations of sonority in ambiguous consonant clusters that would require further exploration and validation by native speaker intuition. We have also gained an insight into the distribution of different syllable structures and syllable nuclei following this approach, which will be useful for comparison with the performance of alternative syllabification systems.

In the future, we plan to compare our system to an onset-maximization-based syllabifier for Serbian in combination with the prescriptive rules to see if we can create an alternative system that will produce outputs consistent with the intuition of native speakers of Serbian. It would be interesting to see a systematic comparison of our current approach and the onset-maximization approach with data gathered from the intuition of contemporary native speakers of Serbian.

We also believe that, while phonological criteria present a basis for syllabification, in the future we will also need to test whether and to what extent approaches based solely on phonological criteria result in syllable boundaries that coincide with morphological boundaries. Our assumption is that phonological rules will need to be amended by morphological criteria to result in syllabification that respects morphological boundaries as well.

In addition to these, the question of the treatment of foreign origin words and transcribed foreign words might be an additional point to consider. As an extension of a syllabifier, a language detection algorithm might be employed to properly segment the former, while the latter might not need special treatment as the process of transcription should in itself contain a degree of phonological adaptation.

Acknowledgment

This research was supported by the Serbian Ministry of Education and Science under the projects Development of Dialogue Systems for Serbian and Other South Slavic Languages (TR-32035) and Languages and Cultures in Time and Space (ON-178002).

Sources and Literature

Literature:

- Barber, Horacio, Marta Vergara, and Manuel Carreiras. 2004. "Syllable-frequency Effects in Visual Word Recognition: Evidence from ERPs." *Neuroreport* 15 (3): 545–48.
- Bradley, Dianne C., Rosa M. Sánchez-Casas, and José E. García-Albea. 2007. "The Status of the Syllable in the Perception of Spanish and English." *Language and Cognitive Processes* 8 (2): 197–233.

- Bigi, Brigitte, and Caterina Petrone. 2014. "A Generic Tool for the Automatic Syllabification of Italian." In *Proceedings of The First Italian Conference on Computational Linguistics, CLiC-it*, 73–77. Pisa: Pisa University Press. http://siti.fileli.unipi.it/projects/clic/proceedings/Proceedings-CLICit-2014.pdf.
- Butt, Matthias. 1992. "Sonority and the Explanation of Syllable Structure." *Linguistische Berichte* 137: 45–67.
- Cholin, Joana, Willem J. M. Levelt, and Niels O. Schiller. 2006. "Effects of Syllable Frequency in Speech Production." *Cognition* 99 (2): 205–35.
- Cholin Joana, and Willem J. M. Levelt. 2009. "Effects of Syllable Preparation and Syllable Frequency in Speech Production: Further Evidence for Syllabic Units at a Post-lexical Level." *Language and Cognitive Processes* 24(5): 662–84.
- Clements, George N. 1990. "The Role of the Sonority Cycle in Core Syllabification." In *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, edited by John Kingston, John and Mary E. Beckman, 282–333. Cambridge: Cambridge University Press.
- Daelemans, Walter, and Antal van den Bosch. 1992. "Generalization Performance of Backpropagation Learning on a Syllabification Task." In *Connectionism and Natural Language Processing: Proceedings of the 3rd Twente Workshop on Language Technology, TWLT3*, 27–38. Enschede: University of Twente, Department of Computer Science. https://pure.uvt.nl/portal/files/760578/generalization.pdf.
- Foley, James. 1972. "Rule Precursors and Phonological Change by Meta-rule." In *Linguistic change and generative theory*, edited by Robert P. Stockwell and Ronald K. S. Macaulay, 96–100. Bloomington: Indiana University Press.
- Goldsmith, John A. 1995. *The Handbook of Phonological Theory*. London: Blackwell Publishers.
- Gvozdanović, Jadranka. 2011. "Phonological Domains." In *Sandhi Phenomena in the Languages of Europe,* edited by Henning Andersen, 27–54. Berlin: Mouton de Gruyter.
- Hankamer, Jorge, and Judith Aissen. 1974. "The Sonority Hierarchy." In *Papers from the Parasession on Natural Phonology*, edited by Anthony Bruck, Robert Allen Fox, and Michael W. La Galy, 131–45. Chicago: Chicago Linguistic Society.
- Hunt, Andrew. 1993. "Recurrent Neural Networks for Syllabification." *Speech Communication* 13 (3–4): 323–32.
- Iacoponi, Luca, and Renata Savy. 2011. "Sylli: Automatic Phonological Syllabification for Italian." In *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association*, 641–44. Florence: International Speech Communication Association. http://eden.rutgers.edu/~li51/php/papers/interspeech2011.pdf.
- Kaplar, Sebastijan, Marija Radojičić, Ivan Obradović, Biljana Lazić, and Ranka Stanković. 2018. "Solution for Quantitative Analysis of Texts in Serbian Based on Syllables." In *ICIST 2018 Proceedings* 2, 315–20. Belgrade: Society for Information Systems and Computer Networks. http://www.eventiotic.com/eventiotic/library/paper/429.
- Kašić, Zorka. 2014. "Opšta lingvistika 2 (Fonologija)." Lecture Materials, Faculty of Philosophy, University of Belgrade.
- Koehler, Klaus J. 1966. "Is the Syllable a Phonological Universal?" *Journal of Linguistics* 2: 207–208.
- Kovač, Aniko, and Maja Marković. 2018. "A Rule-Based Syllabifier for Serbian." In *Proceedings of the Conference on Language Technologies and Digital Humanities 2018*, 140–46. Ljubljana: Ljubljana University Press.
- Ladefoged, Peter, and Keith Johnson. 2014. *A Course in Phonetics*. Belmont: Wadsworth Publishing.
- Ladefoged, Peter. 1982. *A Course in Phonetics*. New York: Harcourt Brace Jovanovich.
- Landsiedel, Christian, Jens Edlund, Florian Eyben, Daniel Neiberg, and Björn Schuller. 2011. "Syllabification of Conversational Speech Using Bidirectional Long-Short-Term Memory Neural Networks." In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5256–9. Prague: IEEE. http://ieeexplore.ieee.org/abstract/document/5947543.

- Marchand, Yannick, Connie R. Adsett, and Robert I. Damper. 2009. "Automatic Syllabification in English: A Comparison of Different Algorithms." *Language and Speech* 52 (1): 1–27.
- Mehler, Jacques, Jean Yves Dommergues, Uli Frauenfelder, and Juan Segui. 1981. "The Syllable's Role in Speech Segmentation." *Journal of Verbal Learning and Verbal Behavior* 20 (3): 298–305.
- Meštrović, Ana, Sanda Martinčić-Ipšić, and Mihaela Matešić. 2015. "Postupak automatskoga slogovanja temeljem načela najvećega pristupa i statistika slogova za hrvatski jezik." *Govor*, 32: 3–34.
- Morelli, Frida. 1999. "The Phonotactics and Phonology of Obstruent Clusters in Optimality Theory." PhD diss., University of Maryland.
- Ohala, John, and Haruko Kawasaki. 1984. "Prosodic Phonology and Phonetics." *Phonology Yearbook*, 1: 113–27.
- Ohala, John. 1990. "The Phonetics and Phonology of Aspects of Assimilation." In *Papers in Laboratory Phonology I*, edited by John Kingston, John and Mary E. Beckman, 258–75. Cambridge: Cambridge University Press.
- Popović, Zoran. 2010. "Taggers Applied on Texts in Serbian." *INFOtheca* 11 (2): 21a–38a.
- Selkirk, Elisabeth O. 1984. "On the Major Class Features and Syllable Theory." In *Language Sound Structure*, edited by Mark Aronoff and Richard T. Oehrle, 107–36. Cambridge: MIT Press.
- Stanojčić, Živojin, and Ljubomir Popović. 2005. *Gramatika srpskoga jezika*. Belgrade: Zavod za udžbenike i nastavna sredstva Beograd.
- Stoianov, Ivelin, John Nerbonne, and Huub Bouma. 1997. "Modelling the Phonotactic Structure of Natural Language Words with Simple Recurrent Networks." In *Computational Linguistics in the Netherlands 1997: Selected Papers from the Eight Clin Meeting*, 77–95. Amsterdam: Rodopi.
- Subotić, Ljiljana, Dejan Sredojević, and Isidora Bjelaković. 2012. *Fonetika i fonologija: Ortoepska i ortografska norma standardnog srpskog jezika*. Novi Sad: Filozofski fakultet Univerziteta u Novom Sadu.
- Utvić, Miloš. 2011. "Annotating the Corpus of Contemporary Serbian." *INFOtheca* 12 (2): 36a–37a.
- Zec, Draga. 2000. "O strukturi sloga u srpskom jeziku." *Južnoslovenski filolog* 56 (1–2): 435–48.

Aniko Kovač, Maja Marković

A MIXED-PRINCIPLE RULE-BASED APPROACH TO THE AUTOMATIC SYLLABIFICATION OF SERBIAN

SUMMARY

In this paper we present a mixed-principle rule-based approach to the automatic syllabification of Serbian based on prescriptive rule descriptions from traditional grammar found in Stanojčić and Popović (2005), extended by rule specifications from Kašić (2014) and Zec (2000), and complemented by a sonority sequencing module based on Selkirk (1984), Subotić et al. (2012), and Zec (2000).

Syllable segmentation plays a role in speech technologies – most notably in the areas of speech recognition and text-to-speech synthesis – at both the segmental and prosodic levels. It is also one of the governing factors in hyphenation, and syllable frequency distribution data is used in psycholinguistic experiments as a covariate. The unavailability of segmented data for Serbian makes a rule-based approach to automatic syllabification the only viable option as there is no data available for training a data-driven neural network model, and the segmentation of large-scale language corpora by trained annotators would be a resource and cost heavy undertaking.

Our goal in this paper is threefold: i) we extend and improve an earlier version of our syllabification algorithm by introducing a sonority sequencing validation module which resolves a number of issues present in the earlier version of our syllabifier, ii) we provide a detailed analysis of the outcomes of the automatic syllabification process in order to address possible theoretical considerations and serve as a basis for the development of future syllabifiers, and iii) we present the statistical data related to the distribution of syllables and their structure in Serbian to be used in psycholinguistic experiments.

The implementation of the existing set of prescriptive rules for the segmentation of words into syllables in Serbian allowed us to gain an insight into problem areas of the rule descriptions, and propose a number of revisions and amendments to the existing rules. The sonority sequencing module revealed the need for an additional onset-length limitation constraint, and pointed out the limitations of sonority in ambiguous consonant clusters – such is the case with continuant fricative phonemes that seem to be able to occupy either the first place in the onset of a syllable or the last place in the coda of a previous syllable – that would require further exploration and validation by native speaker intuition.

The data on the distribution of different syllable structures and syllable nuclei following this approach will be useful for comparison with the performance of alternative syllabification systems. In the future, it would be interesting to see a systematic comparison of our current approach to alternative approaches such as an onset-maximization approach evaluated on segmentation data gathered from the native speakers of Serbian.

Aniko Kovač, Maja Marković

MEŠANI PRISTOP K AVTOMATSKEMU ZLOGOVANJU V SRBŠČINI NA PODLAGI NAČEL IN PRAVIL

POVZETEK

V tem prispevku predstavljamo mešani pristop k avtomatskemu zlogovanju v srbščini na podlagi načel in pravil, ki temelji na opisih predpisnih pravil tradicionalne slovnice (kot jih navajata Stanojčić in Popović 2005), razširjenih z opredelitvami pravil (kot jih navajata Kašić (2014) in Zec (2000)) in dopolnjenih z modulom za zaporedje glede na zvočnost (na podlagi del avtorjev Selkirk 1984; Subotić et al. 2012; Zec 2000).

Členitev na zloge ima pomembno vlogo v govornih tehnologijah – zlasti na področjih prepoznavanja govora in pretvorbe besedila v govor – na segmentalni in prozodični ravni. Je tudi eden od vodilnih dejavnikov pri deljenju besed. Podatki o frekvenčni porazdelitvi zlogov se uporabljajo v psiholingvističnih poskusih kot sočasna spremenljivka. Pristop k avtomatskemu zlogovanju, ki temelji na pravilih, je edina smiselna izbira, saj za srbščino ni na voljo segmentiranih podatkov, iz katerih bi se model nevronske mreže lahko učil. Projekt, pri katerem bi usposobljeni komentatorji razčlenjevali obsežne jezikovne korupse, pa bi bil zelo zahteven in drag.

Naš prispevek ima tri cilje: i) razširiti in izboljšati predhodno različico našega algoritma za zlogovanje z vpeljavo modula za potrjevanje zaporedja glede na zvočnost, ki odpravlja vrsto težav iz predhodne različice našega zlogovalnika; ii) predstaviti podrobno analizo rezultatov avtomatskega postopka zlogovanja, da bi spodbudili morebitne teoretične razmisleke in zagotovili podlago za razvoj prihodnjih zlogovalnikov; in iii) predstaviti statistične podatke, povezane s porazdelitvijo in strukturo zlogov v srbščini, ki jih bo mogoče uporabiti pri psiholingivstičnih poskusih.

Uporaba uveljavljene zbirke predpisnih pravil za členitev besed na zloge v srbščini nam je omogočila, da smo dobili podroben vpogled v težavna področja pri opisih pravil in predlagali vrsto sprememb in popravkov uveljavljenih pravil. Modul za zaporedje glede na zvočnost je razkril potrebo po dodatni omejitvi dolžine vzglasja in izpostavil omejitve zvočnosti pri dvoumnih soglasniških sklopih (na primer priporniki, ki očitno lahko zavzemajo prvo mesto na začetku zloga ali zadnje mesto na koncu predhodnega zloga), ki bi jih bilo treba dodatno raziskati in potrditi s pomočjo intuicije rojenega govorca.

Podatke o porazdelitvi različnih zlogovnih struktur in jeder, pridobljene s tem pristopom, bo mogoče uporabiti za primerjavo z delovanjem drugih sistemov za zlogovanje. Zanimivo bi bilo opraviti sistematično primerjavo našega pristopa z drugimi pristopi, na primer pristopom maksimizacije vzglasja, ovrednotenim na podlagi podatkov o členitvi, pridobljenih od rojenih govorcev srbščine.

1.01 UDC: 003.295:342.537.6:355.012(492)"1940/1945"

Milan M. van Lange,[*] Ralf D. Futselaar[**]

Debating Evil: Using Word Embeddings to Analyse Parliamentary Debates on War Criminals in the Netherlands

IZVLEČEK

RAZPRAVE O ZLU: ANALIZIRANJE PARLAMENTARNIH RAZPRAV O VOJNIH ZLOČINCIH NA NIZOZEMSKEM Z VEKTORSKIMI VLOŽITVAMI BESED

Predstavljamo metodo za raziskovanje sprememb v zgodovinskem diskurzu, pri kateri se uporabljajo obsežni besedilni korpusi in modeli vektorske vložitve besed. Kot študijo primera raziskujemo razprave o kaznovanju vojnih zločincev v nizozemskem parlamentu v obdobju 1935–1975. Predstavili bomo, kako se za sledenje zgodovinskega razvoja parlamentarnega besedišča skozi čas lahko uporabljajo modeli vektorske vložitve besed, ki se učijo z Googlovim algoritmom Word2Vec.

Ključne besede: vojni zločinci, zgodovina kaznovanja, parlamentarna zgodovina, Word2Vec, modeli vektorske vložitve besed

ABSTRACT

We are proposing a method to investigate changes in historical discourse by using large bodies of text and word embedding models. As a case study, we investigate discussions in

[*] **NIOD, Institute for War, Holocaust and Genocide Studies, Herengracht 380, 1016CJ Amsterdam, The Netherlands, m.van.lange@niod.knaw.nl**

[**] **NIOD, Institute for War, Holocaust and Genocide Studies, Herengracht 380, 1016CJ Amsterdam, The Netherlands, r.futselaar@niod.knaw.nl**

Dutch Parliament about the punishment of war criminals in the period 1945–1975. We will demonstrate how word embedding models, trained with Google's Word2Vec algorithm, can be used to trace historical developments in parliamentary vocabulary through time.

Keywords: War Criminals, Penal History, Parliamentary History, Word2Vec, Word Embedding Models

The Case: War Criminals

Soon after German forces in the Netherlands surrendered in May of 1945, the question arose how the hundreds of suspected war criminals and thousands of Nazi collaborators in Dutch custody were to be treated. For the next five decades, this question caused a series of heated political controversies. The debates in Dutch parliament about the punishment, penalty reduction, or release of these people are not only among the longest debates in Dutch parliamentary history, but are generally considered to have been the most emotionally charged (Bootsma and Griensven 2003; Futselaar 2015; Tames 2013).

Discourse and Controversy

In this paper, we use an implementation of word embedding models (WEMs) to analyse parliamentary discussions concerning incarcerated war criminals and Nazi collaborators after the end of the German occupation. At peak, in the summer of 1945, more than a hundred thousand people were incarcerated. They were accused of a variety of crimes, all committed during the occupation of the country: political and military collaboration, war crimes, and (complicity in) genocide. The majority of these prisoners were civilians, whose crimes amounted to little more than membership of national socialist organisations. These people, and other small fry, were released quickly. A small and dwindling number of serious offenders remained in prison, some of them until 1989. After the 1960s, all remaining prisoners were former German officials and officers, whose initial death sentences had been commuted to life in prison. These prisoners became the flashpoint of intense political and media attention. As long as they remained behind bars, plans for their release continued to resurface, and cause political controversy (Piersma 2005; Tames 2013; Futselaar 2015; Grevers 2013).

The main medium of parliamentary communication is spoken language. We aim to demonstrate that a systematic investigation of the verbatim records of the language used in Dutch parliament to discuss these cases can reveal historical change. The results will enable us to track the vocabularies in these discussions through time. We assume that this vocabulary, as we will call it, reflects the changing parliamentary discourse about incarcerated war criminals in Dutch society. We aim to link these developments

in parliamentary vocabulary to actual historical events, developments concerning the post-war dealing with war criminals, and discursive shifts in Dutch society (Olieman et al. 2017). Specifically, we aim to investigate the changing political attitude towards incarcerated war criminals and use our findings to test established notions prevalent in Dutch historiography.

The published proceedings of the two houses of parliament provide us with a dataset comprising of all the words spoken in the plenary sessions. The completeness of the parliamentary dataset allows us to investigate the changing parliamentary vocabulary through time, and in the context of different discussions.

We here focus on two questions directly related to the treatment of these delinquents in the Dutch penal system. The first of these concerns the focus on the identification of the wronged party: did politicians focus on crimes against the Dutch nation as a whole, or against specific groups of individual victims? The second concerns the appropriateness of harsh punishments, specifically whether or not life imprisonment was considered a just alternative for the death penalty. These questions both derive directly from historiography and serve to answer an overarching question: can we assess the validity of traditional scholarship using unsupervised text mining?

Parliamentary Proceedings

In this investigation, we rely entirely on parliamentary proceedings, known in Dutch as the Handelingen der Staten-Generaal. The Handelingen are available in machine-readable form. The minutes of both houses of parliament for the period 1814–1995 were first digitised by the Royal Library of the Netherlands and made available to the public in 2010. The dataset was dramatically improved in the PoliticalMashUp project that ran from 2012 to 2016. This improved and enriched dataset is freely available, on request, from DANS, the Dutch national repository of research data. The dataset consists of a large collection of XML files containing the complete minutes of all the meetings of the lower and upper chambers of parliament, separated by date, speaker, political affiliation, etc. This makes it an excellent corpus for various forms of automated text analysis (Marx et al. 2012).

Word Embedding Models and Historical Research

We investigate the vocabularies used in parliament to discuss a broad category of inmates that could be described as political delinquents, as well as the changes of these vocabularies through time. This is a fairly normal investigation to undertake in traditional historical research - that is to say without computational analyses. Historians typically work by reading the relevant texts. This enables them to use and expand their domain knowledge while processing the data. Although this hermeneutic step is

inevitably part of historical research, this approach has several disadvantages. In this particular case the corpus to be assessed is enormous, making reading and manual encoding of text problematic. More importantly, the traditional research process is highly vulnerable to the biases of the reader/researcher. When studying ethically charged controversies in the relatively recent past, this vulnerability to bias is evidently problematic. People with an interest in recent history and knowledge of the Dutch language almost inevitably hold an opinion on these issues and on the actors in the debate. How do we ensure that our personal political preferences do not influence our reading of the source materials?

Words in Vector Space

A WEM provides a possible solution to these problems. WEMs are techniques to investigate words, and relations between words, in large text corpora. WEMs are based on the calculation of the average distance of unique words to all other unique words in a corpus. The position of each unique word can then be described as a list of numerical values, representing its distance to all other unique words. This list of values is called the 'vector' of the word. In principle, the number of values, also referred to as 'coordinates', or 'dimensions' of the vector, is the same as the number of unique words in the text, minus one. The complete trained corpus, or 'spatial model', is often referred to as a vector space. The method does not prioritize any particular words; the position of each unique word is investigated and given a vector in the model.

The vectors of words within a corpus can be compared. That is to say, the closeness of one vector to another can be calculated. High closeness often reflects a close semantic relationship. Some words with similar vectors are synonyms or near synonyms, or have very similar usages (tea and coffee, for example). Here, we use cosine similarity to calculate the closeness of vectors, although other methods are also feasible.

Since the position of unique words relative to other words is an average calculated on the basis of all occurrences in the text, WEMs are exceptionally effective at investigating relations between relatively frequent words in a sufficiently large text corpus. For historical research, insight in these relations is very useful, and goes far beyond mere closeness. With WEMs we are able to identify associations between words that are not self-evident and would not have been found by traditional means (Schmidt 2015).

Limitations of WEMs

WEMs also have an important downside that is particularly relevant to historical research. Since the training of the model determines the position of a word relative to all other words in that specific corpus, its vector is meaningless in any other

model. Word vectors, hence, can only be compared with other word vectors within the same spatial model. For historians, this means that comparisons between different moments in time are difficult. To make a comparison through time it would be necessary to divide the corpus into subsets representing different periods. For each of these period-specific corpora, a new model, based on a subset of the corpus, needs to be trained. Since vectors of different WEMs are not readily comparable, change through time is difficult to investigate with WEMs. This means that, while WEMs are perfectly adequate tools for fulfilling the first of our aims, investigating vocabularies, they are virtually useless for the second aim, investigating change through time. Since change through time is the core of virtually all historical research (including this investigation), this presents us with a major problem; how can we compare outcomes for different WEMs, for different periods in time?

We have, however, developed a workaround to enable us to use WEMs to investigate changing ways to talk about certain topics through time. We do not directly compare the closeness of vectors within different models, but we calculate relative closeness of vectors for the same terms within different models by using cosine similarity.

Word2Vec

For this investigation, we have used the relatively popular Word2Vec implementation of WEMs to train and analyse word embedding models. Word2Vec was developed by a team of Google engineers and published in 2013. It has been shown to be a particularly effective implementation. This algorithm, however, was developed with a different aim than the one for which we are using it. Initially, Word2Vec was a tool to investigate natural language itself, for example to identify (near) synonyms. In our, historical, investigation, the statistical modelling of language as such is not the objective. Rather than trying to identify linguistic regularities to investigate language, we focus on linguistic irregularities and patterns to identify the influence of political and historical change on the language used in political speech.

For researchers using the R programming language, a package is readily available to analyse texts. This package, created and maintained by Benjamin Schmidt, has been used in this investigation as well (Schmidt 2015, 2017). Our method, however, is in no way dependent on this particular platform and could also be used in Python or any other environment. Neither is the method reliant on the Word2Vec algorithm. It would work broadly in the same way with another implementation of word embeddings. Here, however, we have chosen to use a popular WEM implementation in a relatively user friendly and accessible environment, with the added benefit of using open-source, free software.

Analytical Process

Text analysis with WEMs involves two necessary steps. The first of these, the training of the corpus, creates the spatial model, the WEM itself. The second step is the analysis of the positions of specific words or word clusters within the virtual space of the model.

The corpus of the Handelingen is vast by the standards of historical research (millions of words per year), but not very large for the kind of analysis we are undertaking. For the purpose of WEMs, the size is barely adequate. Therefore we have trained our dataset with a Skip-GramWord2Vec model, which has anecdotally been shown to yield better results on smaller samples (Gelbukh 2015). The vectors of different words can be compared within the model by using cosine similarity. Within a vector space, any two vectors by can be described, by definition, as lying within a horizontal plane. Cosine similarity calculates the angle between these vectors. Perfectly overlapping vectors would result in a cosine similarity of 1, a perfectly opposite relationship -1. In practice, WEMs consist only of positive space, which means that scores fall between 0 (low, or no similarity) and 1 (high, or perfect) similarity (Singhal 2001).

Training the Models

The first step of our workaround is to train two WEMs (more than two is equally feasible), based on two subsets of the corpus (in this case 1945–1955 and 1965–1975). Each of these subsets contains ten years of parliamentary speeches. When using this approach, it is necessary to use relatively similar training corpora, both in terms of size and in terms of language use. For historical research into relatively short periods of parliamentary history, this is not particularly problematic. For reasons of efficiency, we have limited ourselves to unique words that appear at least five times in the corpus and we have limited the number of dimensions of each vector to one hundred. This allows this investigation to be undertaken, and repeated, using fairly normal office grade hardware. We have experimented with more dimensions (several hundreds), but more vectors appear only to be useful with larger corpora. Training WEMs with several hundreds of dimensions also requires far more computational power.

Analysing Word Vectors

Within each spatial model, we have identified the 250 words with the highest cosine similarity to the Dutch terms for 'war criminal' (singular and plural, see Table 1). With these 250 nearest neighbours, we have defined the time specific vocabulary used in the discussion of war criminals. Obviously, these are not the same 250 words in each model. To identify changes in the discussions surrounding our topic,

we calculated the cosine similarity of each of the 250 nearest-neighbour words in each model to two different terms that are present in each of the two corpora. This allows us to compare the position of the vocabulary of the discussion on our topic (war criminals) in relation to, in this case, two stable concepts. The selection of these concepts is crucial for our investigation and for this method. It is here that we translate our research question into a formal, computational inquiry.

For now, we have chosen a two-dimensional implementation of this technique. This is not theoretically necessary, but it allows us to visualize and analyse results more easily in two dimensions. What is important is that concepts used to investigate the relative position of each investigated word are the same in each of the models to be compared. It is also necessary that the concepts are relatively stable through time. Since concepts are represented by words in the corpus itself, words that shift meaning dramatically, such as the English word 'gay', are less suitable than 'cheerful' or 'homosexual', which have not undergone such dramatic change over time.

When discussing concepts, the number of possible words referring to the same concept is often greater than one. Since our investigation focuses on concepts that may be described with multiple words, we need to create a so-called combined vector. We used synonyms and plurals to create a cluster of words with the shared meaning of the concept of interest. This cluster was used as a combined vector in the model by calculating the mean of all the vectors of the cluster words. That is to say that this word set was treated as a single term, resulting in a vector of similar length to a single-word vector. This combined vector allows us to investigate our corpus using all synonyms and near-synonyms of terms as if they were a single term, with a single vector.

Table 1: Word sets used in Debating Evil

Concept	Concept represented by combined vector of the Dutch words:
Death penalty	'doodstraf' and 'doodstraffen'
Life imprisonment	'levenslang', 'levenslange', 'vrijheidsstraf', 'gevangenisstraffen', 'gevangenisstraf', 'opsluiting', and 'hechtenis'
Treason/traitor	'landverrader', 'landverraders', 'verrader', 'verraders', and 'landverraad'
Victim	'slachtoffer' and 'slachtoffers'
War Criminal	'oorlogsmisdadiger' and 'oorlogsmisdadigers'

After selecting two concepts that are present in each of the two corpora, we can calculate the relative similarity of other terms in the corpus to each of them. Although vectors between the two trained WEMs are not comparable, the relative distance to two or more other vectors can be compared very well across several models, provided the underlying concepts are historically stable. When the terms used to estimate the relative position of vocabularies are related and dissimilar, or even perfectly opposite, a historically meaningful analysis becomes viable.

Using two concepts allows us to plot our 'vocabulary', that is the top 250 war-criminal-related words in each of the two periods, in a two-dimensional space. Figure 1 and 2 show the similarity scores of each of the 250 word vocabularies relative to one concept that serves as the y-axis, and another on the x-axis. Each point represents one of the 250 words that form the war-criminal vocabulary for a specific time period. They are plotted based on their cosine similarity score to the combined vector of the concept 'victim' (x) and 'treason' (y) in Figure 1, and to 'life imprisonment' (x) and 'death penalty' (y) in Figure 2. The average scores of all 250 war criminal words on the two dimensions are shown as horizontal and vertical lines. Thus, we have arrived at a visual representation that allows for a comparison of word embedding results for more than one corpus and hence for a comparison through time (in this case, between two distinct historical periods).

Results

Here, we present only two examples using four concepts and two time periods (1945–1955 and 1965–1975). Specifically, we try to identify differences in the way incarcerated war criminals and collaborators were discussed in the immediate aftermath of the Nazi occupation of the Netherlands, and at the height of controversies surrounding the intended release of a number of German war criminals from Dutch prisons - namely Kotälla, Aus der Fünten, and Fischer (Piersma 2005).

Obviously, the discussions in the two periods refer to different groups of perpetrators. In the immediate aftermath of the Nazi occupation the population of inmates was large and diverse, consisting of small-time war profiteers, minor collaborators and their families, but also mass murderers. In the second period, only a handful of elderly foreigners were left, whose crimes were relatively similar and also similarly egregious.

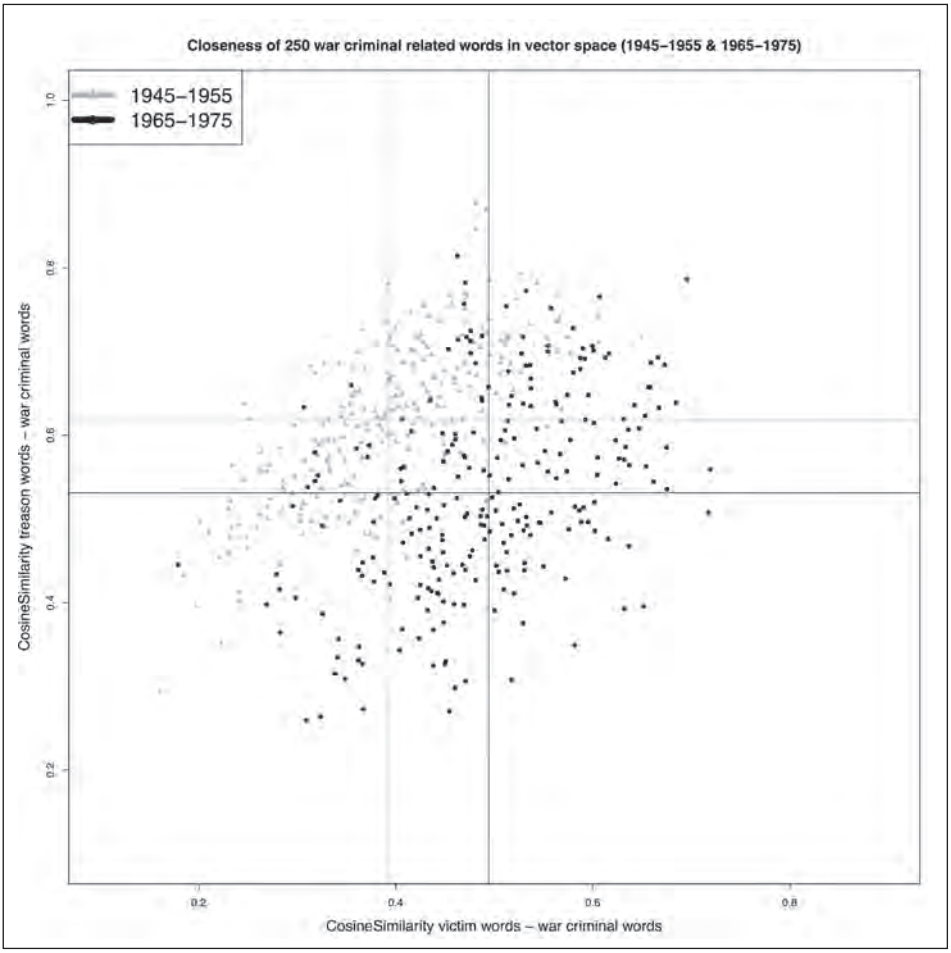
For this investigation, however, our primary aim is not to unearth radically new insights into post-war penal policy in the Netherlands, but to confront the results of an unsupervised, 'distant' reading of parliamentary records to an established historiography. Such a historiography is available for the case at hand; Dutch historians have identified a number of trends in the thinking about political delinquents that (if true) should be reflected in these discussions. Two changes have been identified in particular:

1. A turn in focus from the nature of the crime committed and the person of the perpetrator towards the lasting, psychological damage endured by the victims (Heijden 2012; Haan 1997).
2. A decline in the support, both public and political, for harsh, vengeful punishments, exemplified here in the discussions about the propriety of the death penalty. Although the death penalty was (again) abolished in the 1950s, it remained a point of discussion with regard to war criminals in custody (Futselaar 2015; Smits 2008).

Historical Case

Over the course of three decades, attitudes to incarcerated war criminals, as represented by the vocabularies used to discuss them, changed. In the first period the emphasis lay on crimes against the collective, whereas the focus shifted more towards the plight of individual victims. As can be seen in Figure 1, the initial emphasis on crimes against the nation (treason) in debates about war criminals declined. The average cosine similarity between war-criminal words and treason words (horizontal lines) decreased significantly when we compare 1945–1955 to 1965–1975. At the same time, we observed increased levels of closeness in vector space between war criminal related words to words associated with (individual) victims, as can be seen in Figure 1.

Figure 1: Top 250 war criminal related words 1945–1955 (grey) and 1965–1975 (black) plotted by their cosine similarity to victim (x) and traitor (y) words.



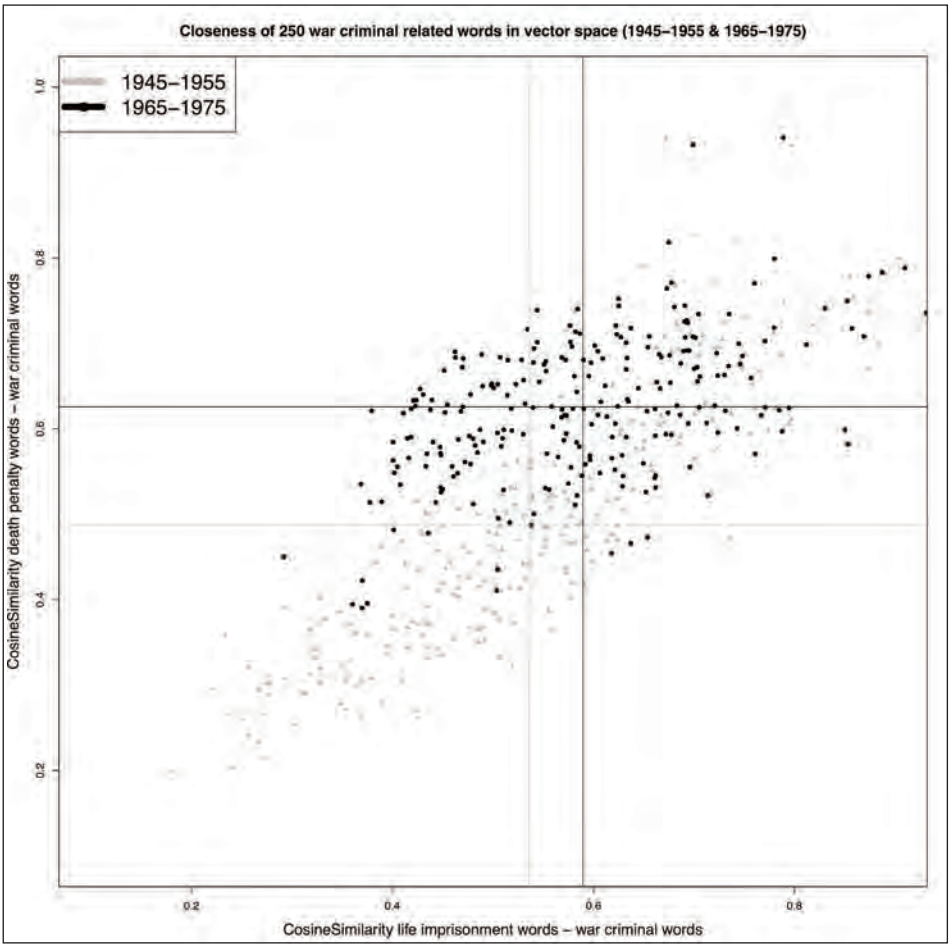
At first glance, this observation is completely in line with the relevant historiography. Several authors have emphasized the sharp rise of interest into the mental health of individual war victims and their families as a decisive factor in policy making and the formation of political opinion. Figure 1 also indicates the observed shift in discourse from focusing on the initial crimes, committed by the war criminals, to the consequences of their deeds for individual people involved (Haan 1997; Heijden 2012; Smits 2008; Withuis 2002).

This development can, however, not be considered a mere discursive change: the observed shifts in parliamentary vocabulary represent actual historical developments in the post-war dealing with war criminals. In the early 1970s, the only war criminals remaining in Dutch prisons were German nationals. Whereas in 1945, main part of the more than hundred thousand incarcerated war criminals were Dutch citizens. Evidently, the accusation of treason was only applicable to the latter group. Hence, if we compare the two periods, it is not surprising that the discursive element of 'treason' decreased in importance in the war criminal vocabulary in Dutch parliamentary debates between 1965 and 1975.

Although the shifts in vocabulary indicate that there was an observable shift in discourse, we have to stress that our analysis also indicates continuity in the parliamentary vocabulary of 1945–1955 and 1965–1975. The scatterplots in Figure 1 indicate a shift, but do not show a complete turn of the parliamentary vocabulary on war criminals. The scatterplots in Figure 1 from both periods show overlap between the nearest neighbours of war criminal related words from 1945–1955 and 1965–1975, scored on closeness to both treason and victim words. We have observed a significant change, or shift. However, we also have to conclude that we did not find a complete turn in vocabulary, as our analysis also indicates continuity and a lasting importance for perpetration and treason in the war criminal debates.

It remains imperative to remain aware of the possible pitfalls of this type of investigation. This is evident in the sharp rise of references to the death penalty in war criminal vocabulary that we observed (see Figure 2). During the second period under scrutiny, capital punishment had long been discontinued in the Netherlands and could not have been discussed as a serious penal option. Closer scrutiny of the data revealed that in many discussions, capital punishment was not advocated, but merely used as a reference point. The war criminals in question had originally been condemned to die, but their punishment had been commuted into life imprisonment. Several members of parliament felt that a pardon would mean that the original verdict (death penalty) would be watered down twice. In these discussions, capital punishment was often referenced, even when its application was not a viable (or even legal) option (Futselaar 2015).

Figure 2: Top 250 war criminal related words 1945–1955 (grey) and 1965–1975 (black) plotted by their cosine similarity to life imprisonment (x) and death sentence words (y).



Conclusion

This paper outlines a method for studying discursive changes in history. We trained WEMs and calculated cosine similarities between two opposite or related concepts for specific periods. This enabled us to compare WEMs for different periods. This opens the door for the use of word embeddings as a tool for historical research, because it enables us to investigate change through time in sufficiently large and consistent historical textual datasets. Parliamentary records are perhaps the best example of such datasets. This method holds considerable promise because parliamentary proceedings and other historical sources are increasingly digitised and made available in machine-readable form.

We have shown how developments in vocabulary can be considered reflective of discursive changes. These changes are related to historical events and developments in the post-war dealing with war criminals in Dutch society. Recent historiography has suggested a dramatic shift away from the crime committed by war criminals and towards the consequences of these deeds for victims and their relatives. We do recognize that victims became more prominent in discussions about war criminals, but this did not diminish the importance of the deed they committed. In other words, the shift is there, but it appears to be far less radical then suggested.

We could also demonstrate that actual historical developments regarding the type of war criminals incarcerated in the Netherlands (from many local convicts, to a handful of foreigners) were reflected by a discursive shift, in which closeness to 'treason' declined. German officials, in the eyes of post-war Dutch parliamentarians, did not commit treason by committing crimes against the Dutch nation.

We have also encountered examples of pitfalls of an overly enthusiastic reliance on word embeddings as an analytical tool. Capital punishment was mentioned particularly frequently in the 1970s, but not because the possibility of executing the war criminals was seriously entertained. Distributional semantics are a powerful new tool for historians, but they do not remove the need for hermeneutic awareness. In this paper, the method is itself the main object of inquiry. We believe we have shown that it possible, feasible, and useful to develop and implement a coherent and widely applicable method for investigating historical change using WEMs.

Discussion

Method Evaluation

For this paper, we have used two corpora, each representing ten years of parliamentary debate to train our WEMs. More interesting, from a research perspective, would be to find out how stable our results are when using smaller, overlapping windows of corpora over time, say with one year steps. It is likely (but not certain) that using more fine-grained windows will reveal similar developments and shifts in language use over time. Repeating the analysis with more data points has the potential to gain more insights in the graduality and the pace of the observed shifts in language used. That said, there is a potential trade-of between detail and precision given that the corpora available to historians are mostly modest in size.

A second ambition is to look more seriously into the distribution of the cosine similarity scores, and the changes in these distributions over time. It will be interesting to measure, visualise, and statistically evaluate these distributions more closely, and to see whether they can be linked to, for example, unanimity and/or homogeneity in parliamentary discussions.

Historical Evaluation

Another remaining ambition is to compare the parliamentary vocabularies used to discuss 'domestic' collaborators and foreign (usually German) war criminals. Furthermore, we also hope to position the war criminal debates in a broader context: how distinct are they from other war related debates, and from other discussions about penal law or criminals in a more general sense? Just as a closer investigation of different categories of perpetrators is viable and useful, different groups of war victims who were discussed in parliamentary debates also license further investigation. These may have included first and second generation victims of wartime violence and persecution, former forced labourers, holocaust survivors and the children of holocaust victims, etc. Given the emphasis on the protection of war victims mentioned above, we are interested to see if there have been changes in the groups emphasized in political debate about the topic.

Acknowledgements

We are grateful to the participants of our Text Mining workshop at the Luxembourg Centre for Contemporary and Digital History (C^2DH) in Esch-sur-Alzette (June 2018), for their comments, input, and criticism. We would also like to thank the participants and organisers of the Language Technologies and Digital Humanities Conference in Ljubljana (September 2018).

Sources and Literature

Datasets and Academic Software:

- Van Lange, Milan. *Debating Evil Repository*. Distributed by Github. https://github.com/MilanvanL/debating_evil.
- Marx, M., J. Van Doornik, A. Nusselder, and L. Buitinck. 2012. "Thematic Collection: PoliticalMashup and Dutch Parliamentary Proceedings 1814–2013." Distributed by *Data Archiving and Networked Services (DANS)*. https://doi.org/10.17026/dans-zg8-9x2v.
- Schmidt, Benjamin. 2017. "Bmschmidt/WordVectors: Tools for Creating and Analyzing Vector-Space Models of Texts Version 2.0 from GitHub." GitHub. Accessed on November 5, 2017. https://rdrr.io/github/bmschmidt/wordVectors/.
- Wickham, Stefan Milton Bache and Hadley. 2014. *Magrittr: A Forward-Pipe Operator for R (version 1.5)*. https://CRAN.R-project.org/package=magrittr.

Literature:

- Bootsma, Peter, and Peter van Griensven. 2003. "'Teleurstelling Is Mijn Opperste Emotie': Vragen over Emotie in de Politiek Aan A.A.M. van Agt." In *Jaarboek Parlementaire Geschiedenis, 2003. Emotie in de Politiek*, edited by Carla van Baalen, Willem Breedveid, Jan Willem Brouwer, Peter van Griensven, Jan Ramakers, and Inke Secker, 121 – 25. Den Haag: SDU Uitgevers.
- Futselaar, Ralf. 2015. *Gevangenissen in oorlogstijd: 1940–1945*. 1st ed. Amsterdam: Boom.
- Gelbukh, Alexander. 2015. *Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14–20, 2015, Proceedings*. Springer.
- Grevers, Helen. 2013. *Van landverraders tot goede vaderlanders: de opsluiting van collaborateurs in Nederland en België, 1944–1950*. Amsterdam: Balans.
- Haan, Ido de. 1997. *Na de ondergang: de herinnering aan de Jodenvervolging in Nederland 1945–1995*. Den Haag: SDU.
- Heijden, Chris van der. 2012. *Dat nooit meer: de nasleep van de Tweede Wereldoorlog in Nederland*. 3rd ed. Amsterdam: Atlas Contact.
- Olieman, Alex, Kaspar Beelen, Milan van Lange, Jaap Kamps, and Maarten Marx. 2017. "Good Applications for Crummy Entity Linkers? The Case of Corpus Selection in Digital Humanities." *CoRR* abs/1708.01162. http://arxiv.org/abs/1708.01162.
- Piersma, Hinke. 2005. *De Drie van Breda: Duitse Oorlogsmisdadigers in Nederlandse Gevangenschap, 1945–1989*. 1st ed. Amsterdam: Balans.
- Schmidt, Benjamin. 2015. "Vector Space Models for the Digital Humanities." Ben's Bookworm Blog. Accessed October 25, 2015. http://bookworm.benschmidt.org/posts/2015-10-25-Word-Embeddings.html.
- Singhal, Amit. 2001. "Modern Information Retrieval: A Brief Overview." *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 24: 9.
- Smits, Hans. 2008. *Strafrechthervormers en hemelbestormers: opkomst en teloorgang van de Coornhert-Liga*. Amsterdam: Aksant.
- Tames, Ismee. 2013. *Doorn in het vlees: foute Nederlanders in de jaren vijftig en zestig*. Erfenissen van Collaboratie. Amsterdam: Balans.
- Withuis, Jolande. 2002. *Erkenning: van oorlogstrauma naar klaagcultuur*. Amsterdam: De Bezige Bij.

Milan M. van Lange, Ralf Futselaar

DEBATING EVIL: USING WORD EMBEDDINGS TO ANALYSE PARLIAMENTARY DEBATES ON WAR CRIMINALS IN THE NETHERLANDS

SUMMARY

This paper presents a case study to investigate the application of text mining techniques in historical research. We demonstrate the usability, advantages, and limitations of distributional semantics when investigating large diachronic historical datasets with word embedding models (WEMs). WEMs are applied to a large digitised and

machine-readable historical dataset, namely the verbatim proceedings of both houses of Dutch parliament for the period 1945–1975.

WEMs are techniques to investigate relations between words in large corpora. WEMs are based on the calculation of the average distance of unique words to all other unique words in a corpus. The position of each unique word can then be described as a list of numerical values, representing its distance to all other words. This list of values is called the 'vector' of the word. These numerical vectors can be compared. That is to say, the closeness of one vector to another can be calculated. High closeness often reflects a close semantic relationship between words. Some words with similar vectors are (near) synonyms or have very similar usages (tea and coffee, for example). For historical research insight in these relations is very useful. It goes far beyond mere closeness. With WEMs we are able to identify associations between words that are not self-evident and would not have been found by traditional means.

The paper uses WEMs to investigate a case study on the vocabulary in parliamentary discussions concerning the punishment, incarceration, and release of Nazi collaborators and war criminals in the Netherlands. We identify changes related to historical events and developments in the post-war dealing with war criminals. Recent historiography on the topic has suggested a dramatic shift away from the crime committed by war criminals and towards the consequences of these deeds for victims and their relatives. We focus on two questions directly related to the treatment of these delinquents in the Dutch penal system. The first of these concerns the focus on the identification of the wronged party: did politicians focus on crimes against the Dutch nation as a whole, or against specific groups of individual victims? The second concerns the appropriateness of harsh punishments, specifically whether or not life imprisonment was considered a just alternative for the death penalty. These questions both derive directly from historiography and serve to answer an overarching question: can we assess the validity of traditional scholarship using text mining?

In the paper we show how victims became more prominent in discussions about war criminals. This did, however, not diminish the importance of the deed they committed. In other words, the shift is there, but it appears to be far less radical then suggested. We also demonstrate that actual historical developments regarding the type of war criminals incarcerated in the Netherlands (from many local convicts in 1945, to a handful of foreigners in the 1970s) were reflected by a discursive shift in the debates. This paper also shows examples of pitfalls of an overly enthusiastic reliance on WEMs as an analytical tool in historical research. Capital punishment was mentioned particularly frequently in the debates of the 1970s, but not because MPs discussed the actual possibility of executing the war criminals.

To conclude: distributional semantics are a powerful new tool for historians, but they do not remove the need for hermeneutic awareness. In this paper, the method is itself the main object of inquiry. We believe we have shown that it possible, feasible, and useful to develop and implement a coherent and widely applicable method for investigating historical change using WEMs. We believe that the outcomes of this

investigation show that WEMs can be a useful and powerful tool in historical research, provided they are used cautiously and with sufficient domain knowledge.

Milan M. van Lange, Ralf Futselaar

RAZPRAVE O ZLU: ANALIZIRANJE PARLAMENTARNIH RAZPRAV O VOJNIH ZLOČINCIH NA NIZOZEMSKEM Z VEKTORSKIMI VLOŽITVAMI BESED

POVZETEK

V prispevku je prikazana študija primera, pri kateri se proučuje uporaba metod za rudarjenje besedil v zgodovinskih raziskavah. Predstavljamo uporabnost, prednosti in omejitve distribucijske semantike pri proučevanju obsežnih diahronih zgodovinskih podatkovnih nizov z modeli vektorske vložitve besed (*word embedding models* – modeli WEM). Modele WEM smo uporabili za analizo obsežnih digitaliziranih in strojno berljivih zgodovinskih podatkovnih nizov, in sicer dobesednih zapisov postopkov v obeh domovih nizozemskega parlamenta v obdobju 1945–1975.

Modeli WEM so metode za proučevanje povezav med besedami v obsežnih korupusih. Temeljijo na izračunu povprečne oddaljenosti edinstvenih besed od vseh drugih edinstvenih besed v korpusu. Položaj vsake edinstvene besede se potem lahko opiše kot seznam numeričnih vrednosti, ki predstavlja njeno oddaljenost od vseh drugih besed. Seznam vrednosti se imenuje "vektor" besede. Te numerične vektorje je mogoče primerjati. To pomeni, da je mogoče izračunati, kako blizu so si posamezni vektorji. Če so si zelo blizu, to pogosto pomen, da so besede tesno semantično povezane. Nekatere besede s podobnimi vektorji so (skoraj) sopomenke ali imajo zelo podobno rabo (na primer čaj in kava). Vpogled v te povezave je zelo koristen za zgodovinske raziskave in presega samo vprašanje bližine. Z modeli WEM lahko prepoznamo povezave med besedami, ki niso očitne in jih ne bi bilo mogoče najti na tradicionalne načine.

V prispevku smo uporabili modele WEM za proučitev študije primera besedišča iz parlamentarnih razprav o kaznovanju, zaporni kazni in izpustitvi nacističnih kolaborantov in vojnih zločincev na Nizozemskem. Ugotavljali smo spremembe, povezane z zgodovinskimi dogodki in dogajanjem v povojni obravnavi vojnih zločincev. V novejšem zgodovinopisju, posvečenem tej tematiki, lahko opazimo precejšen premik od zločinov, ki so jih zagrešili vojnih zločinci, k posledicam teh dejanj za žrtve in njihove sorodnike. Osredotočili smo se na dve vprašanji, ki sta neposredno povezani z obravnavo teh zločincev v nizozemskem sistemu kazenskega pregona. Prvo vprašanje je povezano z osredotočanjem na opredelitev žrtev: ali so se politiki osredotočali na zločine proti nizozemskemu narodu kot celoti ali proti posameznim skupinam

individualnih žrtev? Drugo vprašanje zadeva ustreznost strogih kazni, zlasti ali je dosmrtna zaporna kazen veljala za pravično alternativo smrtni kazni. Obe vprašanji izhajata neposredno iz zgodovinopisja in omogočata odgovor na širše vprašanje: ali lahko presojamo tehtnost tradicionalne znanosti z rudarjenjem besedil?

V prispevku smo pokazali, kako lahko žrtve dobijo pomembnejše mesto v razpravah o vojnih zločincih. S tem pa se ni zmanjšal pomen dejanj, ki so jih zločinci zagrešili. Povedano drugače, premik je mogoče opaziti, vendar se zdi, da je precej manjši od pričakovanega. Pokazali smo tudi, da so se dejanski zgodovinski dogodki, povezani z vojnimi zločinci, ki so bili na Nizozemskem kaznovani z zaporom (od številnih lokalnih obsojencev leta 1945 do nekaj tujcev v sedemdesetih letih 20. stoletja), izrazili v diskurzivnem premiku v razpravah. V prispevku so prikazani tudi primeri različnih pasti, ki jih prinese preveč navdušeno opiranje na modele WEM kot analitično orodje v zgodovinskih raziskavah. Smrtna kazen se je pogosto omenjala predvsem v razpravah v sedemdesetih letih 20. stoletja, vendar ne zato, ker bi poslanci razpravljali o dejanski možnosti usmrtitve vojnih zločincev.

Zaključimo lahko, da je distribucijska semantika koristno novo orodje za zgodovinarje, vendar to ne pomeni, da hermenevtična zavest ni več potrebna. V tem prispevku je glavni predmet proučevanja sama metoda. Menimo, da smo dokazali, da je mogoče, izvedljivo in koristno razviti in uporabljati usklajeno ter za široko rabo primerno metodo za proučevanje zgodovinskih sprememb z modeli WEM. Verjamemo, da rezultati te raziskave dokazujejo, da so modeli WEM lahko koristno in uporabno orodje v zgodovinskih raziskavah, če jih uporabljamo previdno in z ustreznim znanjem.

1.01 UDC: 004.774-026.11

Andrej Pančur*

Sustainability of Digital Editions: Static Websites of the History of Slovenia – SIstory Portal

IZVLEČEK

TRAJNOST DIGITALNH IZDAJ: STATIČNE SPLETNE STRANI PORTALA ZGODOVINA SLOVENIJE – SISTORY

Prispevek izhaja iz stališča, da je pri digitalnih izdajah potrebno poskrbeti za čim bolj celovito digitalno trajnost tako podatkov kot prezentacij, funkcionalnosti in programske kode. To je velik izziv predvsem za manjše digitalno humanistične projekte z omejenim financiranjem, ki ne omogoča dolgoročnega vzdrževanja tehnično zahtevnih digitalnih izdaj. Kot alternativno rešitev so v prispevku predstavljene rešitve, ki jih v zadnjih letih ponuja hiter razvoj statičnih spletnih strani. Digitalne izdaje, ki temeljijo na TEI, so s pomočjo osnovnih XML (XSLT) in spletnih tehnologij (HTML, CSS, JavaScript) kot statične spletne strani uspešno vključene v repozitorij portala SIstory. Vse statične spletne strani imajo tudi možnost dinamičnega prikazovanja vsebine.

Ključne besede: digitalne izdaje, digitalno kuratorstvo, TEI, XSLT, statične spletne strani

ABSTRACT

The contribution is based on the position that, with regard to digital editions, the highest possible degree of digital sustainability of data, presentations, functionalities, and programme code should be ensured. This represents a significant challenge, especially in case of smaller digital humanities projects with limited financing, which does not allow for the long-term maintenance of technically-demanding digital editions. The alternative solutions facilitated by the swift development of static websites in the recent years are presented in the

* **Institute of Contemporary History, Kongresni trg 1, SI-1000 Ljubljana, andrej.pancur@inz.si**

contribution. Digital editions based on the TEI have been successfully included in the SIstory portal repository as static websites, employing basic XML (XSLT) and web technologies (HTML, CSS, JavaScript). All the static websites also have the possibility of displaying dynamic content.

Keywords: digital editions, digital curation, TEI, XSLT, static website

Introduction

In digital humanities, the awareness of the importance of digital sustainability and permanent preservation of digital sources has been present for a long time (Schaffner and Erway 2014, 7). The research data of an individual project usually outlives the project in the context of which it has been collected, organised, and published. Therefore it is very important to ensure a high-quality and sustainable storage of digital data even after the project itself has been concluded.

In the recent years, the technical aspects of research data management and long-term archiving (metadata, archive formats, preservation media, and documentation) have been the subject of intensive discussions. Only lately, however, have we begun to realise that the preservation of data in accordance with the specific requirements of various scientific disciplines is almost more important for the high-quality management and reuse of this data (Moeller et al. 2018). While in the natural and social sciences the data from measurements and questionnaires is typically used, in the humanities the use of cultural objects like manuscripts, texts, pictures, and recordings is predominant. Moreover, researchers in humanities will usually additionally process, visualise, tag, link, and interpret digital cultural objects (DHd-AG Datenzentren 2017, 7).

Such data processing is particularly important in case of digital editions, which are a crucial part of digital humanities (Andorfer et al. 2016). Naturally, digital scholarly editions mostly consist of the research in the context of which different transcriptions, indications, analyses, explanations, etc., are produced. Such research data in particular should therefore be available to the research community in the long term and under open access conditions (Robinson 2016). In the case of digital editions, the encoded text is the most crucial long-term result of the project. The display of information is vital as well, as it represents the outlook of the project group on this information in the context of a certain application. However, it is not that every such outlook is unique in any way or even the only one possible. Instead, this information can be displayed in a variety of ways (Turska et al. 2016). With each new interpretation, the number of other potential user interfaces even increases. Each such presentation is thus a new research result that deserves long-term storage as well.

Therefore, research results in humanities consist not only of research data, but also of the presentation environment and the applications that enable data interpretation,

searching, filtering, browsing, and linking (DHd-AG Datenzentren 2017, 7). If we only stored research data, the initial presentation would be lost forever, even though the presentation represents an integral part of any digital edition (Fechner 2018). At the same time, we should not forget that the programming code used for the creation of digital editions is an integral part of the scientific argumentation as well, just like the digital editions (Andrews and Zundert 2016).

Sustainable storage of digital editions therefore represents a particularly significant challenge. Moreover, digital editions can be very different from each other in terms of their contents, appearance, and functionality. They mostly result from specific research projects with relatively limited financial and human resources at their disposal. As the project group members come from the field of humanities, they often lack the suitable technical expertise, which is why they mostly need to rely on external contractors when it comes to technical development. Furthermore, digital editions depend on the very swift development of online technologies and standards (Andorfer et al. 2016).

As the number of digital editions increases rapidly, the challenges involved in the sustainable storage of digital editions will only become greater in the future (Fechner 2018). In case of smaller digital humanities projects with limited financing, which does not allow for the long-term maintenance of technically-demanding digital editions, this represents a significant challenge and will continue to do so. In the continuation, I will present alternative solutions offered by the rapid development of static websites. In the recent years, static websites have become one of the main online development trends. It appears that this trend will also persist in the future (Williams 2019). In the present contribution, I will present the experience gained by generating static websites for the digital editions in the context of the activities of the Research Infrastructure of Slovenian Historiography, which, among other tasks, also manages the History of Slovenia – SIstory web portal.[1] In this regard I will restrict my article solely to the static websites generated from XML files, encoded in accordance with the Text Encoding Initiative Guidelines (TEI) (TEI Consortium 2019). In digital humanities, the TEI Guidelines are the *de facto* standard for text encoding, used by many different humanities projects and studies (Romary et al. 2017, 5).

In the chapter *Modern Static Websites*, I will first present the main advantages and disadvantages of this type of websites. In our case, we have decided to upgrade the basic XSLT Stylesheets of the TEI Consortium. In the *SIstory TEI Profile* chapter, I will present generic upgrade of the TEI Stylesheets. In the chapter *Configuring and Upgrading the SIstory TEI Profile* I will outline the project-specific options for upgrading this profile. In both these chapters, I will also discuss the various options of adding dynamic contents to static websites. In the chapter *Publishing Digital Editions* I will outline how these static websites can be made available to the public, in particular by their inclusion in the SIstory portal's digital repository. In the *Conclusion*, I will also mention a few more general findings.

1 "Research Infrastructure of Slovenian Historiography," *History of Slovenia – SIstory*, accessed April 15, 2019, http://www.sistory.si/publikacije/?menuBottom=2.

Modern Static Websites

All websites used to be static at first, which is why all of the digital editions in the field of digital humanities were initially created as static HTML websites. This was also true in case of the Slovenian scholarly digital editions (Ogrin and Erjavec 2009),[2] which have introduced the paradigm of digital editions in Slovenia (Ogrin 2005). The creators of these digital editions soon encountered certain shortcomings of static websites. In particular, they missed the option of carrying out structured text searches, adaptable URL query string parameters, and dynamic web content association. In the case of newer digital editions, they therefore opted for the Fedora Commons platform (Erjavec et al. 2011).

By that point, the internet had been, for a long time already, dominated by dynamic websites that had successfully replaced the outdated static websites, where the contents could only be altered by the developers directly editing the HTML code. By means of content management systems (e.g. the very popular WordPress, Drupal, and Joomla), dynamic websites have finally made it possible for technically unskilled users to start publishing on the internet.

The contents of dynamic websites are stored in databases. The server does not construct the contents until the user demands that a website be displayed, adapted to the demands of the user. A suitable programming language is used to communicate with the server. The biggest problem of such dynamic websites is that its technical solutions are often more complicated than the actual needs of their users.

Modern static websites, however, have been created as an answer to the problems exhibited by dynamic websites. Unlike the latter, static websites do not employ databases and server-side programming languages, but are simply a collection of HTML, CSS, and JavaScript files. Static websites therefore enjoy numerous advantages in comparison with dynamic websites (Rinaldi 2015):

- efficiency: as static websites do not require any databases or server-side processing, they are not in danger of becoming slow;
- hosting: because static websites do not rely on a server-side programming language, their hosting is simple and cheap. There are even free options, for example the GitHub Pages service;
- security: static websites do not require any databases or server-side programming languages that hackers could breach. Therefore such sites are safe until the files they consist of are stored securely;
- maintenance: as static websites do not rely on any databases, server-side programming languages, or content management systems, their maintenance is extremely simple;
- versioning: since static websites consist exclusively of text files, all of their versions can be quite simply stored in version control systems like Git.

2 *Scholarly Digital Editions of Slovenian Literature, eZISS*, accessed April 15, 2019, http://nl.ijs.si/e-zrc/index-en.html.

These reasons are particularly important to ensure the sustainability of digital editions. The use of standard formats like TIFF and JPEG for digital photographs, HTML and XML for texts, and so on, ensures that the digital editions created will remain readable and useful for a long time to come (Rosselli Del Turco 2016). Consequently, this paradigm started to be emphasised in other similar projects in the field of digital humanities as well (Viglianti 2017; Daengeli and Zumsteg 2017; Diaz 2018).

These reasons, however, are less convincing in case we expect digital editions to contain user-generated contents as well. Therefore, static websites are not appropriate for all digital editions in the field of digital humanities, as such solutions will often fail to satisfy the needs of the creators and users. On the other hand, countless digital projects do not call for very complex content and its display. In such cases the existing solutions provided by static websites can be more than satisfactory, especially because modern static websites do not completely lack the option of adding dynamic contents. In reality, static websites have only experienced their renaissance with the appearance of various services and programming solutions that allowed such websites to include dynamic contents.

Modern static websites are no longer coded manually, but are instead generated by employing static website generators. Nowadays, the selection of such generators is extremely broad. One of the most popular is Jekyll,[3] which is also used in the creation of GitHub pages. Thus its use has also spread to humanities (Visconti 2016). Static website generators assume that the users will write the contents using text formatting syntax like Markdown markup language, which is very popular among developers.[4] These formats can then be converted to HTML sites with a website generator and then published online. However, the Markdown syntax is very deficient and only allows for basic content publishing. As such, it is inappropriate for the tagging of complex humanities texts. Consequently, humanities texts are most often encoded with Extensible Markup Language (XML). Furthermore, XSLT (Extensible Stylesheet Language for Transformation) is used as a tool for XML conversion. Together, these are the key technologies employed by digital humanities (Flanders et al. 2016). As the use of XSLT transformations is often very similar to static site generator conversions, we can describe XSLT as a "modern, efficient static site generator" as well (Kraetke and Imsieke 2016).

SIstory TEI Profile

For many years, the TEI Consortium has been regularly maintaining and updating the XSL Stylesheets, which can be used to generate, on the basis of TEI documents, not only (X)HTML websites, but also many other formats, including LaTeX, XSL-FO, EPUB, DOCX, and ODT. These XSL stylesheets are freely available from

3 *Jekyll • Simple, blog-aware, static sites*, accessed April 15, 2019, https://jekyllrb.com/.

4 *Daring Fireball: Markdown*, accessed April 15, 2019, https://daringfireball.net/projects/markdown/.

the GitHub repository and regularly updated in accordance with the new versions of the TEI Guidelines.[5] Not only is the relevant written documentation very good, but the programming code comments are exemplary as well. XSLT stylesheets are also used, among other things, to generate the static website for each version of the TEI Guidelines.[6]

Most importantly, by means of custom profiles, the XSLT stylesheets of the TEI Consortium allow for very flexible adaptations to different project requirements. In fact, the XSL Stylesheets for TEI have been written with the intention of being as adaptable as possible. Numerous parameters exist that can be configured according to preferences. The stylesheets contains many variables and templates, which can be adapted to specific requirements. The authors of the code even thought of empty (hook) templates, to which custom contents and XSLT programming code may be added. I have made use of all these options when writing the SIstory profile for the XSLT stylesheets of the TEI Consortium. (Pančur 2019a)

Initially, I based the creation of these profiles on the needs of the Research Infrastructure of Slovenian Historiography for flexible and prompt publication of our technical documentation online. In the context of the Research Infrastructure, my colleagues and I are managing the History of Slovenia – SIstory portal, which also contains a repository and digital library. Therefore we have decided to include these digital editions into the existing infrastructure as intensively as possible. Until 2016, the static websites of these digital editions had been stored on an additional www2 server of the SIstory portal,[7] while the digital library itself had only stored the metadata about the digital editions and links to these static sites. After the upgrade of the SIstory portal in 2016, we could start storing the HTML and all other files related to these digital editions directly in the repository and the digital library.

Due to the desire to maximize the integration of digital editions into the SIstory portal, I also tried to bring the external appearance of digital editions as close as possible to the user interface of the portal. As an example, Figure 1 shows a snapshot of the home page of the portal between the years 2012 and 2016, and in Figure 2, the user interface of the digital edition of 2014.

5 *TEI XSL Stylesheets*, accessed April 15, 2019, https://github.com/TEIC/Stylesheets.

6 "P5: Guidelines for Electronic Text Encoding and Interchange," *TEI: Text Encoding Initiative*, accessed April 15, 2019, https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html.

7 *www2.SIstory.si*, accessed April 15, 2019, http://www2.sistory.si/.

Figure 1: Home page of the History of Slovenia – SIstory portal of 2016

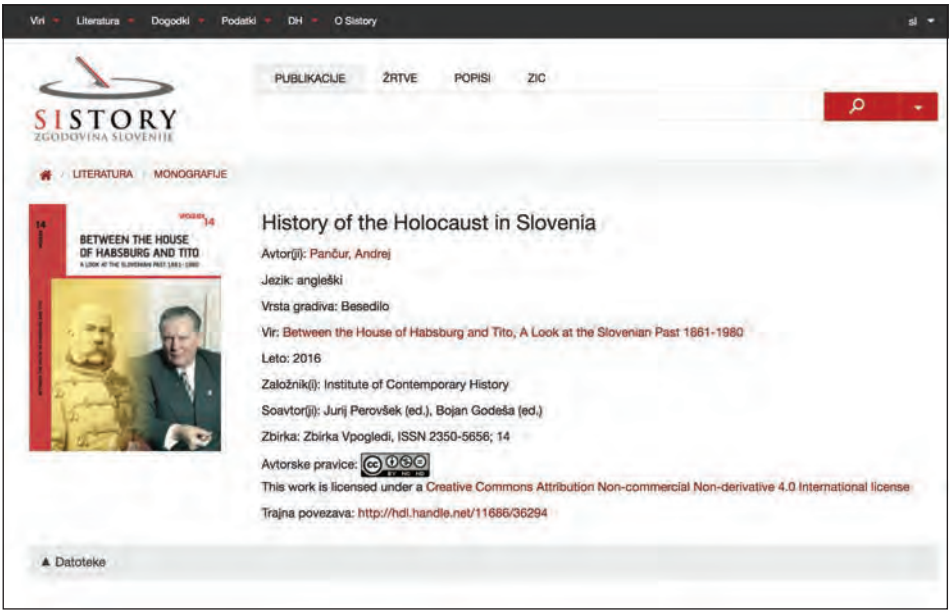
Source: Spletni arhiv Narodne in univerzitetne knjižnice, accessed April 10, 2018, http://nukrobi2.nuk.uni-lj.si:8080/wayback/20160225143401/http://www.sistory.si/.

Figure 2: The 2014 digital edition user interface

Source: (Gašparič 2014), accessed April 10, 2018, http://www2.sistory.si/publikacije/monografije/Gasparic_Parlamentaria1/ch01.html.

Even though the colour scheme is identical and the layout of the logo, the search bar, main top navigation menu, and the contents are very closely modelled after the SIstory portal, the user interfaces are nevertheless not the same. At the time, the user interface of the portal was still based on the old HTML 4 technology, but I had already started to use responsive website design and HTML 5 for the digital editions. In this regard, I decided to use the responsive front-end framework ZURB Foundation.[8] I keep my adaptations as well as CSS and JS additions in the GitHub repository. (Pančur 2019b) As the use of this framework turned out to be extremely useful, we also included it in the new SIstory portal in 2016. Subsequently I also adapted the appearance of the digital editions to the new portal design (compare Figures 3 and 4).

Figure 3: Top navigation menu, search bar, and metadata page of the SIstory portal



Source: (Pančur 2016).

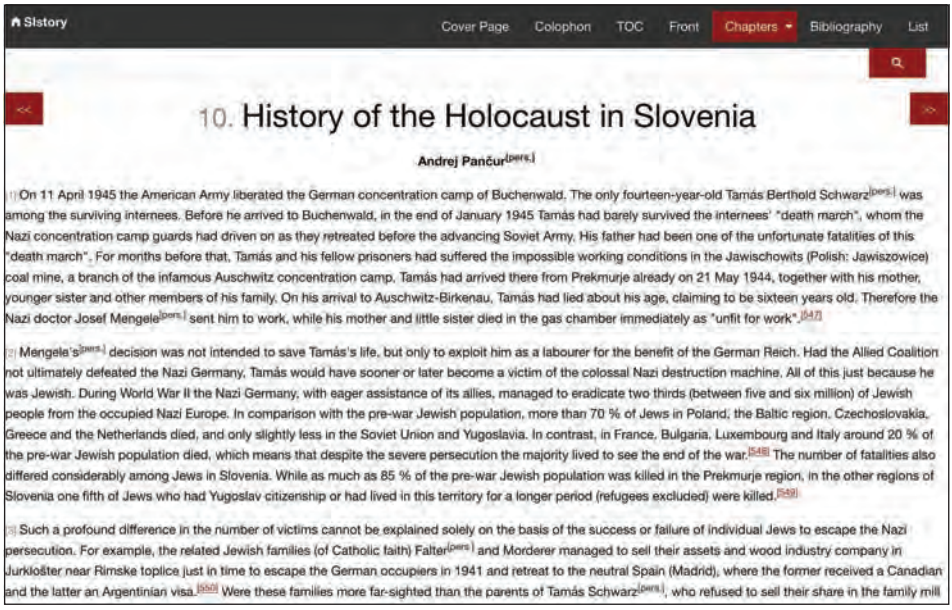
Apart from the originally envisioned technical documentation, we soon also started to publish other sorts of publications – in particular monographs, collections of scientific texts, and magazines – online in the HTML format. Therefore I reconfigured the SIstory TEI profile with the aim of facilitating the publication of these sorts of digital editions. The profile allows for the transformations of:

– individual TEI documents;

8 *Foundation: The most advanced responsive front-end framework in the world*, accessed April 15, 2019, https://foundation.zurb.com/.

– several TEI documents from a shared TEI corpus. In this case, each TEI document needs to be converted separately. The TEI corpus itself and its <*teiHeader*> need to be converted separately, as in this manner a common cover, colophon, and tables of contents are generated.

Figure 4: The 2016 digital edition user interface



Source: (Pančur 2016), accessed April 15, 2019, http://www.sistory.si/cdn/publikacije/36001-37000/36294/ch10.html.

The digital edition's main navigation menu is located at the very top of the web page, as horizontal navigation with a drop-down menu. The structure of this navigation reflects the structure, sections, and divisions of the individual TEI documents. In the continuation I will briefly outline the possible content sections of the navigation as well as the TEI document. In practice, no TEI document contains every single one of these sections. Instead, the authors of TEI documents can use and arrange them completely in accordance with their needs.

The central part of the content is always contained within the <*body*> element. The main content must be contained within a single or several <*div*> elements with the obligatory attribute *@xml:id*. Each <*div*> element represents its own division of the content or chapter. Therefore the navigation bar's single drop-down menu displays all of the <*div*> divisions contained within the <*body*> element. A variety of contents, encoded in the relevant TEI document within the <*front*> and <*back*> elements, may also be accessible before and after this part of the drop-down menu. Figure 5 thus illustrates all of these main content sections.

Figure 5: The main content sections of a TEI document

```
<text>
   <front>
      <titlePage>
         <docTitle>
            <titlePart>Title of digial edition</titlePart>
         </docTitle>
         <docAuthor>Author of digital edition</docAuthor>
         <docEdition>Document edition</docEdition>
         <docImprint>
            <pubPlace>Publication place</pubPlace>
            <docDate>Publication date</docDate>
         </docImprint>
         <graphic url="url_of_cover_page_image.jpg"/>
      </titlePage>
      <div type="preface" xml:id="prf-01">
         <!-- Introductory chapter -->
      </div>
   </front>
   <body>
      <div type="chapter" xml:id="ch01">
         <!-- Chapter with main content -->
      </div>
   </body>
   <back>
      <div type="bibliogr" xml:id="bibl01">
         <!-- Bibliography -->
      </div>
      <div type="appendix" xml:id="app01">
         <!-- Appendix -->
      </div>
      <div type="summary" xml:id="sum01">
         <!-- Summary -->
      </div>
   </back>
</text>
```

Only <*titlePage*> is obligatory, because it is converted to the default start page (index.html) and, as such, accessible through the navigation bar – at the very top, as the Title Page. The <*front*> element may contain one or several <*div*> elements, which represent the introductory chapters section in the navigation. The <*back*> element includes three possible content sections (bibliographies, annexes, summaries), which is why they must always be assigned the appropriate *@type* attribute. Each of these sections can consist of one or more chapters. In most cases, the conversion of the content of these divisions is based on the standard XSLT stylesheets of the TEI Consortium, which I have only partly adapted to the needs of our own digital editions. I have written the transformations for the generated <*divGen*> divisions from scratch.

All of them have been included in the SIstory TEI profile. These generated divisions can be included in the <*front*> (Figure 6) or the <*back*> element (Figure 7), and each of the <*divGen*> elements must include a <*head*> with an arbitrary division title. These titles are then included in the digital edition's navigation.

Figure 6: The list of all possible generated <*divGen*> divisions, contained in the <*front*> element

```
<front>
   <divGen type="cip" xml:id="cip">
      <head>Colophon</head>
   </divGen>
   <divGen type="teiHeader" xml:id="teiHeader">
      <head>TEI header</head>
   </divGen>
   <divGen type="toc" xml:id="id-toc">
      <head>Table of contents</head>
   </divGen>
   <divGen type="toc" xml:id="id-images">
      <head>List of images</head>
   </divGen>
   <divGen type="toc" xml:id="id-charts">
      <head>List of charts</head>
   </divGen>
   <divGen type="toc" xml:id="id-tables">
      <head>List of tables</head>
   </divGen>
   <divGen type="toc" xml:id="id-titleAuthor">
      <head>Table of contents where the name
         of the author is also displayed</head>
   </divGen>
   <divGen type="toc" xml:id="id-titleType">
      <head>Simplified table of contents</head>
   </divGen>
   <divGen type="search" xml:id="search">
      <head>Search</head>
   </divGen>
</front>
```

Unlike the aforementioned <*div*> elements, where the use of *@xml:id* identifiers is merely recommended (the HTML files that contain these divisions are named after these identifiers), in case of generated divisions they are obligatory and also have a semantic meaning that is of key importance for their conversion. The *@type* attribute defines the main category, which is particularly highlighted in the horizontal navigation. The *@xml:id* attribute more precisely defines the subcategory, shown in the navigation drop-down menu. The most extensive category is the Table of Contents (TOC) group, which, apart from the various tables of the contents of chapters and

subchapters, also contains a list of tables, figures, and charts. In reality, the list of charts is merely a separate group of list of figures (<*figure*>), which includes figures with the *@type* attribute and *chart* value.

The <*back*> element involves only a single category of generated divisions that includes various lists of persons, places, and organisations. The generated divisions include all of the persons mentioned in the TEI document, encoded with the <*persName*> element, all places encoded with <*placeName*>, or all organisations encoded with <*orgName*>. All of the named entities, encoded in this manner, must also be assigned the *@ref* attribute, in order to refer to the appropriate canonical element in the list of entities (<*listPerson*> for persons, <*listOrg*> for organisations, and <*listPlace*> for places) in the TEI header (<*teiHeader*>). The <*placeName*> element's *@ref* attribute may also contain a reference to the GeoNames[9] or DBpedia[10] URI, where the SIstory profile can process the geographical coordinates and display them in the list of places.

Figure 7: The list of all possible elements for automatically generated text division <*divGen*>, contained in the <*front*> element

```
<back>
   <divGen type="index" xml:id="id-persons">
      <head>List of persons</head>
   </divGen>
   <divGen type="index" xml:id="id-places">
      <head>List of places</head>
   </divGen>
   <divGen type="index" xml:id="id-organizations">
      <head>List of organizations</head>
   </divGen>
</back>
```

As it is also possible to use the SIstory profile to convert the TEI documents from the TEI corpus, the <*divGen*> elements from the various TEI documents cannot possess the same *@xml:id* identifiers. Therefore the subcategories of the generated divisions are specified in such a manner that the subcategory's identifier is stated after the final hyphen of this identifier's value (see Figures 6 and 7, where the *id* before the hyphen in *@xml:id* attribute defines the arbitrary identifier, while the subcategory is stated after the hyphen).

The SIstory profile also allows for the display of dynamic contents. The Tipue Search engine is included as a basic functionality.[11] It can be included with a generated division (<*divGen*>) of the *search* type in the <*front*> element. Tipue Search is an open source jQuery plugin, which can be relatively easily integrated even in static sites. In

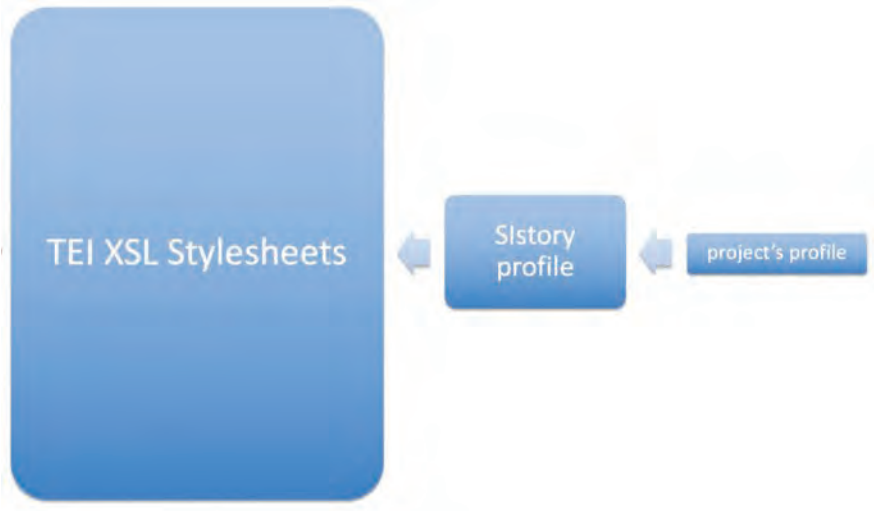
9 *GeoNames*, accessed April 15, 2019, http://www.geonames.org/.
10 *DBpedia*, accessed April 15, 2019, http://wiki.dbpedia.org/.
11 *Tipue Search*, accessed April 15, http://www.tipue.com/search/.

the graphical user interface, the search bar is located immediately below the bottom navigation, while the element <*divGen*> generates a *search.html* web page that includes a dynamic display of search results. The content of the TEI document is indexed, as a JavaScript object (JSON), in the file *tipuesearch_content.js*, which needs to be located in the same folder as the *search.html* file. Content indexation takes place at the level of paragraphs (<*p*>), lists (<*list*>), tables (<*table*>), figures (<*figure*>), and all other possible TEI elements, which are direct child elements of the text division <*div*>. Therefore, all of these elements must include a *@xml:id* attribute for unique identifier. Lists are the only exception: when they do not possess the *@xml:id* attribute, whereas their child elements do, then the latter are indexed.

Configuring and Upgrading the SIstory TEI Profile

Much like the main XSL Stylesheets of the TEI Consortium, the SIstory profile has been created to allow for its adaptation to the requirements of any individual project. To this end, it includes a few original parameters of the TEI Consortium's XSLT stylesheets which affect the default stylesheet output, to which I have added a few new SIstory parameters. All of these parameters can be set up anew for each conversion, but it is more appropriate that new project profiles be created for each individual project. The conversion usually proceeds in the following manner: the project's custom profile imports the SIstory profile, which, in turn, imports the TEI XSLT transformations, and adds overrides (see Figure 8).

Figure 8: Chained XSLT conversions with additional profiles



For example, during conversion, the default SIstory profile thus expects every chapter or the first <*div*> text division to become a separate HTML web page. In this case, navigation through forward and back buttons is added to the web pages automatically. Unlike the original TEI transformations, this navigation also includes the <*divGen*> generated divisions. However, by changing the *splitLevel* parameter (originally a parameter included in the TEI conversions), it is possible to specify that subchapters also become separate HTML web pages. The forward/back and up/down navigation between the web pages has now been appropriately adapted. The current SIstory profile only supports a depth of three text divisions.

The *documentationLanguage* parameter may currently be used to specify the Slovenian, English, or Serbian navigation (in the Latin or Cyrillic script). By adding new translations to the *myi18n.xml* document, it is possible to further expand this localisation. The localisation of the Tipue Search engine has been suitably taken care of as well.

The SIstory profile also allows for the parallel display of the texts' various language versions. In this case, all of the main <*div*> text divisions and <*divGen*> generated divisions must contain *@xml:lang* attributes with the appropriate language code as well as *@corresp* attributes pointing at all the other language versions of the text in question (see Figure 9). Simultaneously, the *languages-locale* parameter must be set to the value *true*, while the *languages-locale-primary* parameter must specify the language code of the starting *index.html* file.

Figure 9: Localization and language setting in TEI document

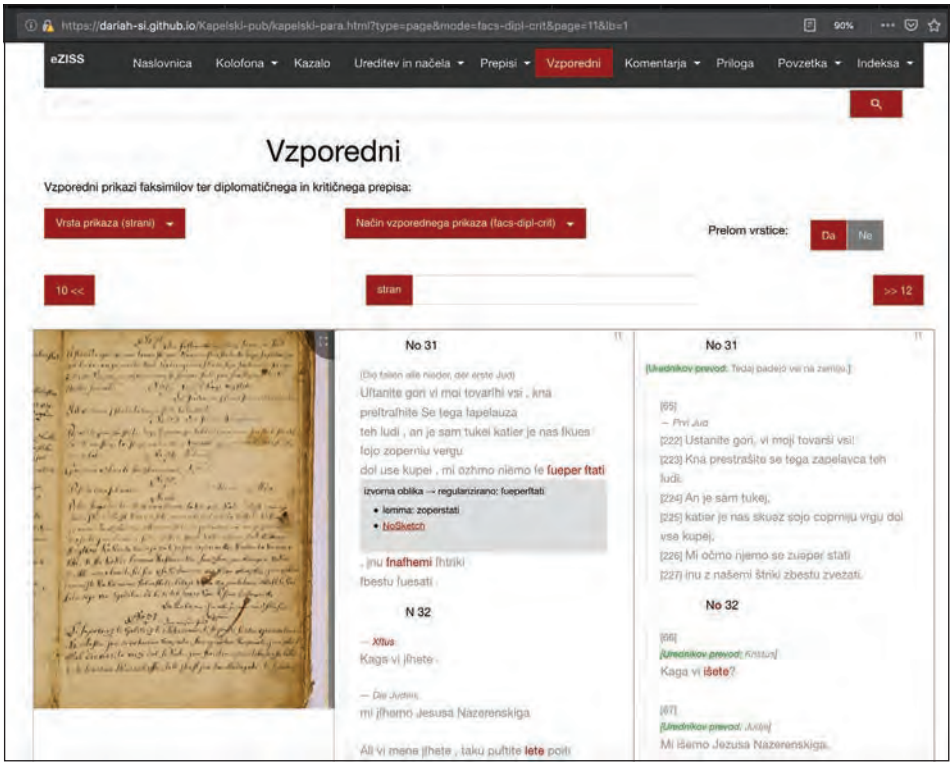
```
<body>
   <div type="chapter" xml:id="ch01"  xml:lang="sl"
      corresp="#en-ch01">
      <!-- Slovenian chapter with main content -->
   </div>
   <div type="chapter" xml:id="en-ch01" xml:lang="en"
      corresp="#ch01">
      <!-- English chapter with main content -->
   </div>
</body>
```

The display of the TEI document's metadata from the <*teiHeader*> element is similarly adaptable. This transformation can be initially specified by including the generic division (<*divGen*>), whose *@type* attribute value should be set to *teiHeader* (see Figure 6). The entire content of the <*teiHeader*> element is converted to <*dl*> (definition list HTML element), where <*dt*> (description term element) defines the name of the TEI element as well as the names and attribute values (element [attribute = value | attribute = value]), while <*dd*> (definition description element) defines the text contents of the TEI element. Of course, the definitions are appropriately nested. With additional parameters, this transformation can be configured in such a way as to

display the descriptive names of elements and attributes in the English or Slovenian language instead of their names.

Apart from this simple SIstory profile configuration, any additional XSLT transformation that can be completely adapted to the needs of an individual digital edition can be included during the conversion of a project. Simultaneously, by using various JavaScript libraries and plugins as well as web applications, it is also possible to enable additional dynamic content display. For example, in the case of the SIstory portal's digital editions, I have successfully used DataTables[12] to display large quantities of tabled data, Highcharts[13] for charts, Google Maps for maps, and ImageViewer[14] and Viewer.js for images.[15] These are merely examples: there are alternatives, and every year many new possibilities emerge.

Figure 10: The simultaneous display of facsimiles, diplomatic transcription, and critical transcription of the *Kapelski pasijon* passion play.



Source: Kapelski pasijon, GitHub pages, https://dariah-si.github.io/Kapelski-pub/

12 *DataTables: Table plug-in for jQuery*, accessed April 15, 2019, https://datatables.net/.
13 *Highcharts*, accessed April 15, 2019, https://www.highcharts.com/products/highcharts/.
14 *ImageViewer*, accessed April 15, 2019, http://ignitersworld.com/lab/imageViewer.html.
15 Viewer.js, JavaScript image viewer, accessed April 15, 2019, https://fengyuanchen.github.io/viewerjs/.

Simultaneously, in 2017, with the publication of Saxon-JS[16], the possibilities of dynamically displaying the contents of XML documents in static web pages have even improved. Saxon-JS is an XSLT 3.0 run-time written in pure JavaScript. It could contribute to XSLT once again becoming a client-side technology that works in a browser (Lumley et al. 2017). For digital editions, I have thus started to successfully use an Saxon extension function *ixsl:query-params*, which parses the query parameters of the HTML page URI. In the case of the *Kapelski pasijon* (The Železna Kapla Passion Play) digital edition, I have thus created and used the following parameters to generate a dynamic parallel display of facsimiles as well as the diplomatic and critical transcription: *type*, *mode*, *page*, and *lb* (line break). These parameters have allowed me to construct a dynamic display of extremely complex contents (Figure 10), which can still be optionally upgraded in the future digital editions.

Publishing Digital Editions

The default SIstory profile transformation generates all the HTML, JS, and any other files in a single folder. As the digital editions generated in this manner consist solely of static web pages, they can also be used on personal computers. In this manner it is possible to effectively test the digital editions even before publishing them online, where we can swiftly and simply publish them on any accessible servers. Additionally, the GitHub repository web pages are a free option that can also ensure an efficient version control.

However, the main purpose of SIstory profiles is to include digital editions directly into the SIstory portal's repository and its digital library. Thus we can efficiently store all of the digital editions' files by adding persistent Handle System identifiers and checksums for all the relevant files, as well as flexibly organise digital editions as one or several digital objects with one or several intellectual entities. Each intellectual entity has its own Handle identifier and metadata. It can include several files or none at all. The files belonging to an individual intellectual entity are located in the same folder. The path to this folder also includes the suffix of the Handle persistent identifier, which is, in the case of the SIstory portal, always a numerical value (e.g., for the suffix 555, the relative path would be */cdn/publikacije/1-1000/555/file*). Therefore, the SIstory XSLT profile must know the values of these identifiers in advance. Thus we can precisely determine, even in advance, whether the entire contents of a digital edition should be contained in a single intellectual entity of the SIstory portal, or whether various digital edition files should be included in various intellectual entities. These identifiers can be recorded among the rest of the metadata in <*teiHeader*>, within the <*publicationStmt*> element, as a value of one or more <*idno*> elements. This element

16 "Saxon-JS," *Saxonica*, accessed April 15, 2019, http://www.saxonica.com/saxon-js/index.xml.

requires that the value of the *@type* attribute be specified as *sistory* or *si4*, while the *@corresp* attribute should point at all the appropriate <*div*> and <*divGen*> divisions whose content will be included in the intellectual entity with this identifier.

The SIstory XSLT profile is open source and available in the GitHub repository. (Pančur 2019a) Another GitHub repository also contains all of the digital editions currently kept on the SIstory portal. The project upgrades of the SIstory XSLT profile for each of these editions are available as well. (Pančur 2018) I regularly expand and maintain the SIstory profile in accordance with the changes of the TEI XSLT stylesheets.

Conclusion

There are several advantages to a digital editions infrastructure organised in this manner:

- using format that is most common in digital humanities: TEI XML (Neuefeind 2019, 221);
- using a single XML technology (XSLT) for various sorts of digital editions, which enjoys a wide support in the TEI community;
- the possibility of simply including JavaScript libraries and plugins;
- flexibly adding dynamic contents with Saxon-JS;
- in comparison with other technologies (dynamic sites), static sites ensure a relative sustainability and simple maintenance of digital editions;
- using Git version control to store the various versions of digital editions together with the software used to generate static websites;
- open access to the complete digital editions code in the GitHub and GitLab software development platforms;
- the possibility of sharing digital editions on the GitHub Pages and GitLab Pages, and, last but not least, the possibility of including them in the History of Slovenia – SIstory portal.

Acknowledgements

The work presented in this paper was supported by the Slovenian historiography research infrastructure (I0- 0013), and the Slovenian ESFRI infrastructures DARIAH-SI which are financially supported by the Slovenian Research Agency.

Sources and Literature

Datasets and Academic Software:

- Pančur, Andrej. 2018. *Electronic publishing on SIstory.* Distributed by GitHub. https://github.com/SIstory/publications.
- Pančur, Andrej. 2019a. *SIstory TEI Stylesheets.* Distributed by GitHub. https://github.com/SIstory/Stylesheets.
- Pančur, Andrej. 2019b. *SIstory: additional CSS and JS.* Distributed by GitHub. https://github.com/SIstory/themes.

Literature:

- Andorfer, Peter, Matej Ďurčo, Thomas Stäcker, Christian Thomas, Vera Hildenbrandt, Hubert Stigler, Sibylle Söring, and Lukas Rosenthaler. 2016. "Nachhaltigkeit technischer Lösungen für digitale Editionen: Eine kritische Evaluation bestehender Frameworks und Workflows von und für Praktiker_innen." In *DHd 2016: Modellierung – Vernetzung – Visualisierun: Die Digital Humanities als fächerübergreifendes Forschungsparadigma: Konferenzabstracts*, 36–39. Universität Leipzig. http://www.dhd2016.de/.
- Andrews, Tara, and Joris van Zundert. 2016. "What Are You Trying to Say? The Interface as an Integral Element of Argument." In *Digital Scholarly Editions as Interfaces, International Symposium at the University of Graz, Austria*, 31–32. Graz: Centre for Information Modelling – Austrian Centre for Digital Humanities. https://static.uni-graz.at/fileadmin/gewi-zentren/Informationsmodellierung/PDF/dse-interfaces_BoA21092016.pdf.
- Daengeli, Peter, and Simon Zumsteg. 2017. "Hermann Burgers Lokalbericht: Hybrid-Edition mit digitalem Schwerpunkt." In *DHd 2017: Digitale Nachhaltigkeit: Konferenzabstracts*, 151–55. Universität Bern. http://www.dhd2017.ch/.
- DHd-AG Datenzentren. 2017. *Geisteswissenschaftliche Datenzentren im deutschsprachigen Raum: Grundsatzpapier zur Sicherung der langfristigen Verfügbarkeit von Forschungsdaten.* Hamburg. DOI: 10.5281/zenodo.1134760.
- Erjavec, Tomaž, Jan Jona Javoršek, Matija Ogrin, and Petra Vide Ogrin. 2011. "Od biografskega leksikona do znanstvenokritične izdaje: vprašanje trajnosti elektronskih besedil." *Knjižnica* 55, No. 1: 103–14. https://knjiznica.zbds-zveza.si/knjiznica/article/view/6004.
- Diaz, Chris. 2018. "Using Static Site Generators for Scholarly Publications and Open Educational Resources." *Code4Lib Journal*, No. 44. https://journal.code4lib.org/articles/1386.
- Fechner, Martin. 2018. "Eine nachhaltige Präsentationsschicht für digitale Editionen." In *DHd 2018: Kritik der digitalen Vernunft: Konferenzabstracts*, edited by Georg Vogeler, 203–7. Universität zu Köln. http://dhd2018.uni-koeln.de/.
- Flanders, Julia, Syd Bauman, and Sarah Connell. 2016. "XSLT: Transforming our XML data." In *Doing Digital Humanities: Practice, Training, Research*, edited by C. Crompton, R. J. Lane and R. Siemens, 255–72. Oxon and New York: Routledge.
- Gašparič, Jure. 2014. *Slovenski parlament: Politično zgodovinski pregled od začetka prvega do konca šestega mandata (1992–2014).* Ljubljana: Inštitut za novejšo zgodovino. http://hdl.handle.net/11686/26950.
- Kraetke, Martin, and Gerrit Imsieke. 2016. "XSLT as a Modern, Powerful Static Website Generator: Publishing Hogrefe's Clinical Handbook of Psychotropic Drugs as a Web App." In *Proceedings of XML in, Web Out: International Symposium on sub rosa XML*, Balisage Series on Markup Technologies, vol. 18. https://doi.org/10.4242/BalisageVol18.Kraetke02.

- Lumley, John, Debbie Lockett, and Michael Kay. 2017. "Compiling XSLT3, in the Browser, in Itself." In *Proceedings of Balisage: The Markup Conference 2017*, Balisage Series on Markup Technologies, vol. 19. https://doi.org/10.4242/BalisageVol19.Lumley01.
- Moeller, Katrin, Matej Ďurčo, Barbara Ebert, Marina Lemaire, Lukas Rosenthaler, Patrick Sahle, Urlike Wuttke, and Jörg Wettlaufer. 2018. Die "Summe geisteswissenschaftlicher Methoden? Fachspezifisches Datenmanagement als Voraussetzung zukunftsorientierten Forschens." In *DHd 2018: Kritik der digitalen Vernunft: Konferenzabstracts*, edited by Georg Vogeler, 89–93. Universität zu Köln. http://dhd2018.uni-koeln.de/.
- Neuefeind, Claes, Philip Schildkamp, and Brigitte Mathiak. 2019. "Technologienutzung im Kontext Digitaler Edition – eine Landschaftsvermessung." In *DHd 2019: Digital Humanities: multimedial & multimodal. Konferenzabstracts*, 219–22. Universität Mainz, Universität Frankfurt. https://dhd2019.org/.
- Ogrin, Matija, and Tomaž Erjavec. 2009. "Ekdotika in tehnologija: Elektronske znanstvenokritične izdaje slovenskega slovstva." *Jezik in slovstvo* 54, No. 6 (2009): 57–72. http://www.dlib.si/?URN=URN:NBN:SI:doc-BOC8BANS.
- Ogrin, Matija, ed. 2005. *Znanstvene razprave in elektronski mediji: razprave*. Ljubljana: Založba ZRC, ZRC SAZU. http://nl.ijs.si/e-zrc/bib/eziss-knjiga.pdf.
- Pančur, Andrej. 2016. "History of the Holocaust in Slovenia." In *Between the House of Habsburg and Tito: A Look at the Slovenian Past*, edited by Jurij Perovšek and Bojan Godeša. Ljubljana: Inštitut za novejšo zgodovino. http://hdl.handle.net/11686/36294.
- Rinaldi, Brian. 2015. *Static Site Generators: Modern Tools for Static Website Development.* Sebastopol, CA: O'Reilly Media.
- Robinson, Peter. 2016. "Why Interfaces Do Not and Should Not Matter for Scholarly Digital Editions." In *Digital Scholarly Editions as Interfaces, International Symposium at the University of Graz, Austria*, 29–30. Centre for Information Modelling – Austrian Centre for Digital Humanities.
- https://static.uni-graz.at/fileadmin/gewizentren/Informationsmodellierung/PDF/dse-interfaces_BoA21092016.pdf.
- Rosselli Del Turco, Roberto. 2016. "The Battle We Forgot to Fight: Should We Make a Case for Digital Editions?" In *Digital Scholarly Editing: Theories and Practices*, edited by Matthew James Driscoll and Elena Pierazzo, 19–238. Cambridge: Open Book Publishers. http://dx.doi.org/10.11647/OBP.0095.
- Romary, Laurent, Piotr Banski, Jack Bowers, Emiliano Degl'innocenti, Matej Ďurčo, Roberta Giacomi, Klaus Illmayer, Adeline Joffres, Fahad Khan, Mohamed Khemakhem, et al. 2017. *Report on Standardization (draft)*. [Technical report] 4.2 Inria. https://hal.inria.fr/hal-01560563.
- Schaffner, Jennifer and Ricky Erway. 2014. *Does Every Research Library Need a Digital Humanities Center?* Dublin, Ohio: OCLC Research. https://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-digital-humanities-center-2014.pdf.
- TEI Consortium, ed. 2019. *TEI P5: Guidelines for Electronic Text Encoding and Interchange 3.5.0.* TEI Consortium. http://www.tei-c.org/Guidelines/P5/.
- Turska, Magdalena, James Cummings, and Sebastian Rahtz. 2016. "Challenging the Myth of Presentation in Digital Editions." *Journal of the Text Encoding Initiative*, No. 9. DOI: 10.4000/jtei.1453.
- Viglianti, Raffaele. 2017. "Your Own Shelley-Godwin Archive: An off-line strategy for an on-line publication (poster)." In *TEI 2017 Victoria*. https://hcmc.uvic.ca/tei2017/abstracts/t_126_viglianti_shelleygodwin.html.
- Visconti, Amanda. 2016. "Building a Static Website with Jekyll and GitHub Pages." *The Programming Historian, 5*. https://programminghistorian.org/lessons/building-static-sites-with-jekyll-github-pages.
- Williams, Martin. 2019. *Web Development Trends 2019* (blog). March 14, 2019. Accessed April 12, 2019. https://www.keycdn.com/blog/web-development-trends-2019.

Andrej Pančur

SUSTAINABILITY OF DIGITAL EDITIONS: STATIC WEBSITES OF THE HISTORY OF SLOVENIA – SISTORY PORTAL

SUMMARY

The contribution is based on the position that, with regard to digital editions, the highest possible degree of digital sustainability of data, presentations, functionalities, and programme code should be ensured. This represents a significant challenge, especially in case of smaller digital humanities projects with limited financing, which does not allow for the long-term maintenance of technically-demanding digital editions. The alternative solutions facilitated by the swift development of static web pages in the recent years are presented in the contribution.

Static websites enjoy numerous advantages in comparison with dynamic websites: efficiency, hosting, security, maintenance, and versioning. These reasons are particularly important to ensure the sustainability of digital editions. These reasons, however, are less convincing in case we expect digital editions to contain user-generated contents as well. Therefore, static websites are not appropriate for all digital editions in the field of digital humanities. On the other hand, countless digital projects do not call for very complex content and its display. In such cases the existing solutions provided by static websites can be more than satisfactory, especially because modern static websites do not completely lack the option of adding dynamic contents. Modern static websites are generated by employing static website generators. Humanities texts are most often encoded with Extensible Markup Language (XML). Extensible Stylesheet Language for Transformation (XSLT) is used as a tool for XML conversion: also in static websites. Digital editions based on the TEI have been successfully included in the SIstory portal repository as static web pages, employing basic XML (XSLT) and web technologies (HTML, CSS, JavaScript). All the static web pages also have the possibility of displaying dynamic content.

In the case of SIstory portal, we have decided to upgrade the basic XSLT Stylesheets of the TEI Consortium. In the *SIstory TEI Profile* chapter, I will present generic upgrade of the TEI Stylesheets. In the chapter *Configuring and Upgrading the SIstory TEI Profile* I will outline the project-specific options for upgrading this profile. In both these chapters, I will also discuss the various options of adding dynamic contents to static websites. In the chapter *Publishing Digital Editions* I will outline how these static websites can be made available to the public, in particular by their inclusion in the SIstory portal's digital repository. In the *Conclusion*, I will also mention a few more general findings.

There are several advantages to a digital editions infrastructure organised in this manner: using format that is most common in digital humanities (TEI XML); using

a single XML technology (XSLT) for various sorts of digital editions, which enjoys a wide support in the TEI community; the possibility of simply including JavaScript libraries and plugins; flexibly adding dynamic contents with Saxon-JS; in comparison with other technologies (dynamic sites), static sites ensure a relative sustainability and simple maintenance of digital editions; using Git version control to store the various versions of digital editions together with the software used to generate static websites; open access to the complete digital editions code in the GitHub and GitLab software development platforms; the possibility of sharing digital editions on the GitHub Pages and GitLab Pages, and, last but not least, the possibility of including them in the History of Slovenia – SIstory portal.

Andrej Pančur

TRAJNOST DIGITALNH IZDAJ: STATIČNE SPLETNE STRANI PORTALA ZGODOVINA SLOVENIJE – SISTORY

POVZETEK

Prispevek izhaja iz stališča, da je pri digitalnih izdajah potrebno poskrbeti za čim bolj celovito digitalno trajnost tako podatkov kot prezentacij, funkcionalnosti in programske kode. To je velik izziv predvsem za manjše digitalno humanistične projekte z omejenim financiranjem, ki ne omogoča dolgoročnega vzdrževanja tehnično zahtevnih digitalnih izdaj. Kot alternativno rešitev so v prispevku predstavljene rešitve, ki jih v zadnjih letih ponuja hiter razvoj statičnih spletnih strani.

Statične spletne strani imajo v primerjavi s dinamičnimi številne prednosti: zmogljivost, gostovanje, varnost, vzdrževanje in kontrola verzij. Ti razlogi so zlasti pomembni zaradi trajnosti digitalnih izdaj. Vendar so ti razlogi manj prepričljivi, če glede digitalnih izdaj pričakujemo, da bodo vsebovale tudi uporabniško generirano vsebino. Zato statične spletne strani niso primerne za vse digitalne izdaje s področja digitalne humanistike. Po drugi strani pa je zelo veliko digitalnih projektov, kjer vsebina in njen prikaz nista tako zelo zahtevni. V teh primerih bi bile obstoječe rešitve, ki jih prinašajo statične spletne strani, več kot zadovoljive, predvsem zaradi tega, ker moderne statične strani niso povsem brez možnosti dodajanja dinamičnih vsebin. Moderne statične spletne strani generiramo s pomočjo generatorjev statičnih spletnih strani. Besedila v humanistiki večinoma kodiramo z XML označevalnim jezikom. XSLT pa uporabljamo kot orodje za pretvorbo XML: tudi v statične spletne strain. Digitalne izdaje, ki temeljijo na TEI, so s pomočjo osnovnih XML (XSLT) in spletnih tehnologij (HTML, CSS, JavaScript) kot statične spletne strani uspešno vključene v repozitorij portala SIstory. Vse statične spletne strani imajo tudi možnost dinamičnega prikazovanja vsebine.

V primeru portala SIstory smo se odločili za nadgradnjo osnovnih pretvorb XSLT konzorcija TEI. V poglavju SIstory TEI profil bom predstavil svojo generično nadgradnjo pretvorb XSLT konzorcija TEI. V poglavju Konfiguracija in nadgradnja SIstory profila bom nato predstavil projektno specifične možnosti nadgradnje tega profila. V obeh teh poglavjih bom predstavil še različne možnosti dodajanja dinamične vsebine statičnim spletnim stranem. V poglavju Publiciranje digitalnih izdaj bom omenil, kako te statične spletne strani damo na razpolago javnosti, predvsem z vključitvijo v digitalni repozitorij portala SIstory. V Sklepu naposled dodam še nekaj pomembnejših splošnih ugotovitev.

Tako vzpostavljena infrastruktura za digitalne izdaje ima več prednosti: uporaba podatkov, ki so v digitalni humanistiki najbolj razširjeni (TEI-XML); uporaba enotne XML tehnologije (XSLT) za različne vrste digitalnih izdaj, ki ima široko podoro v TEI skupnosti; možnost enostavnega vključevanja JavaScript knjižnic in vtičnikov; fleksibilno dodajanje dinamične vsebine s Saxon-JS; statične spletne strani zagotavljajo v primerjavi z ostalimi tehnologijami (dinamične spletne strani) relativno trajnost digitalnih izdaj ter relativno enostavno vzdrževanje; uporaba Git kontrole verzij za shranjevanje različnih izdaj digitalnih izdaj, skupaj s programsko opremo, ki smo jo uporabili pri generiranju statičnih spletnih strani; odprti dostop do celotne kode digitalnih izdaj v platformah za razvoj programske opreme GitHub in GitLab; možnost gostovanja digitalnih izdaj v GitHub Pages in GitLab Pages in nenazadnje možnost vključitve v portal Zgodovina Slovenije – SIstory.

1.01 UDC: 003.295:572+316.7

Ajda Pretnar*, Dan Podjed**

Data Mining Workspace Sensors: A New Approach to Anthropology

IZVLEČEK

PODATKOVNO RUDARJENJE SENZORJEV V DELOVNEM OKOLJU: NOV PRISTOP K ANTROPOLOGIJI

Antropologija po nepotrebnem zaostaja pri vključevanju računskih metod v raziskave, čeprav te postajajo vse bolj priljubljene v družboslovju in humanistiki. Tudi v antropologiji namreč uspešno združujemo kvantitativne in kvalitativne metode, še posebej kadar prehajamo med njimi. V prispevku predlagamo nov metodološki pristop in opišemo, kako smo uporabili kvantitativne metode in podatkovno analitiko v etnografskem raziskovalnem delu. Metodologijo prikažemo na primeru analize senzorskih podatkov ene od fakultetnih stavb Univerze v Ljubljani, kjer smo opazovali prakse in vedenje zaposlenih med delovnim časom in ugotavljali, kako upravljajo s stavbo in bivalnim okoljem. Za raziskovanje smo na primeru t.i. »pametne stavbe« uporabili krožne mešane metode, ki prepletajo podatkovno analitiko (kvantitativni pristop) z etnografijo (kvalitativni pristop), ter sočasno empirično identificirali glavne prednosti nove antropološke metodologije.

Ključne besede: računska antropologija, senzorski podatki, podatkovna etnografija, krožne mešane metode

ABSTRACT

While social sciences and humanities are increasingly including computational methods in their research, anthropology seems to be lagging behind. But it does not have to be so. Anthropology is able to merge quantitative and qualitative methods successfully, especially when traversing between the two. In the following contribution, we propose a new

* **Laboratory of Bioinformatics, Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, SI-1000 Ljubljana, ajda.pretnar@fri.uni-lj.si**

** **Institute of Slovenian EthnologyResearch Centre of the Slovenian Academy of Sciences and Arts, Novi trg 2, SI-1000 Ljubljanadan.podjed@zrc-sazu.si**

methodological approach and describe how to engage quantitative methods and data analysis to support ethnographic research. We showcase this methodology with the analysis of sensor data from a University of Ljubljana's faculty building, where we observed human practices and behaviours of employees during working hours and analysed how they interact with the building and their environment. We applied the proposed circular mixed methods approach that combines data analysis (quantitative approach) with ethnography (qualitative approach) on an example of a "smart building" and empirically identified the main benefits of the new anthropological methodology.

Keywords: computational anthropology, sensor data, data ethnography, circular mixed methods

Introduction

Social sciences and humanities are rapidly adopting computational approaches and software tools, resulting in an emerging field of digital humanities (Klein and Gold 2016) and computational social sciences (Conte et al. 2012). Among these is anthropology, which is particularly suitable for traversing between quantitative and qualitative methods. Anthropologists study and analyse human habits, practices, behaviours and cultures, with a particular focus on participant observation and long-term fieldwork as a methodological cornerstone of the discipline. With an increasing availability of data coming from social networks and wearable devices among other sources (Miller et al. 2016; Gershenfeld and Vasseur 2014), anthropologists can easier than ever dive into data analysis and study humans and their societies, subcultures and cultures quantitatively as well as qualitatively.

With this contribution, we tentatively place anthropology in the field of digital humanities,[1] mostly because the suggested approach is multidisciplinary and by analogy similar to the shifts between distant and close reading (Jänicke et al. 2015) in literary studies. Just like distant reading can offer an abstract (over)view of the corpus, quantitative analyses can give a researcher a broad understanding of the population she is investigating. And just like distant reading needs close reading to understand the style, themes, and subtle meanings of a literary work, so does data analysis need an ethnographic approach to contextualize the information and extract subtle meanings of individual human experience.

As Pink et al. (2017) suggest, there is value in investigating everyday data that reveal what is ordinary, what extraordinary and how to contextualize the two. In this contribution we expand the idea by employing the *circular mixed methods* approach that combines qualitative research from anthropology and quantitative analysis from data mining. We consider the mixed methods (Creswell and Clark 2007; Teddlie and

1 Anthropology is considered a part of the humanities in the Slovenian academic tradition, while elsewhere it is placed under the umbrella of social sciences. In reality, it probably lies at the intersection of both.

Tashakkori 2009) as an integrative research that merges data collection, methods of research and philosophical issues from both quantitative and qualitative research paradigms into a singular framework (Johnson et al. 2007). We also stress the need for a circular research design, where we traverse between methods to continually verify and enhance knowledge. Circularity gives research flexibility and enables shifting perspectives in response to new information.

Our study, which provides the basis for this article, began in October 2017 and includes 14 offices at one of the University of Ljubljana's faculty buildings. The building is equipped with automation systems and sensors measuring large amounts of data related mostly to the building's energy performance and thermal comfort. We retrieved measurements from approximately 20 sensors from the SCADA monitoring system for the year 2016 and extrapolated behavioural patterns for different rooms and, more generally, room types through data visualization and exploratory analysis. The analysis showed specific patterns emerging in several rooms; we noticed there were some definite outliers in terms of working hours and room interaction.

We used computational methods to gauge new perspectives on human behaviour and invoke potentially relevant hypotheses. Data analysis provided several distinct patterns of behaviour and defined the baseline for workspace use. However, this approach was unable to provide us with a context for the data. Quantitative methods can easily answer the 'what', 'where' and 'when' type of questions, but struggle with the 'why'. At that stage, we employed anthropological fieldwork and ethnography as the main methods of anthropology. We conducted interviews with room occupants to explain what the uncovered patterns mean and why people behave the way they do.

The main purpose of our study was to demonstrate how anthropologists can use statistics and data visualisation to establish the essential facts of the observed phenomena and how the traditional anthropological methods, which have not significantly changed since the early 20th century, when Malinowski (2002 [1922]) carried out his ground-breaking ethnographic research at the Trobriand Islands, can be complemented and upgraded by data analysis. We call this a circular mixed methods approach, where circular implies continual traversing between qualitative and quantitative methods, between fieldwork and data analysis. Our research applies the proposed methodology to sensor data obtained from a smart building and with a combination of data mining and ethnographic field establishes both a wide and deep understanding of human behaviour in a workplace setting.

While inclusion of domain experts is already a postulate in machine learning and data mining, the opposite, the inclusion of machine learning and data mining in anthropology, is fairly new and lacks sufficient practical application. Concurrently, few anthropologists and even social scientists and humanists in general are included in the development of AI solutions and data analysis, even when the data is strictly coming from a social domain (Skeem and Lowenkamp 2016; Lum and Isaac 2016; see also Pretnar and Robnik-Šikonja 2019). Moreover, the plot twist in anthropology comes from the fact that anthropologists do not act as domain experts explaining the

data, but as channels and interpreters for the people to explain the data they produced themselves. In anthropology, authority does not come from the researcher, but from the researched – the group of people that are the source of data and information.

Development of Computational Anthropology

While digital humanities became a full-fledged field in the last couple of decades (Hockey 2004), anthropology seems to be lagging behind. Some authors suggest anthropology should be more concerned with digital as an object of analysis rather than as a tool (Svensson 2010). However, there have been several attempts to include computational methods and quantitative analyses into anthropological research. Already in the 1960s, anthropologists looked at using computers for the organisation of anthropological data and field notes (Kuzara et al. 1966; Podolefsky and McCarty 1983) and started using computers for social network analysis (Mitchell 1974). Progress in text analysis, coding facts, and comparative studies in linguistics (Dobbert et al. 1984; White and Truex 1988) followed suit.

However, only lately has there been a significant computational breakthrough in the discipline. Digital anthropology turned disciplinary attention to the analysis of online worlds, virtual identities, and human relationships with technology. For example, Bell (2006) gave a cultural interpretation of the use of ICTs in South and Southeast Asia, Boellstorff (2015) investigated online worlds in the Second Life, Nardi (2010) explored gaming behaviour of the World of Warcraft, and Bonilla and Rosa (2015) described how to use hashtags for ethnographic research. Moreover, a discussion has been opened on what does 'big data' mean for social sciences and how to ethically address its retrieval and analysis (boyd and Crawford 2012; Mittelstadt et al. 2016).

There was a discussion on the methodological front as well. Anderson et al. (2009) argue for a method that combines the ethos of ethnography with database mining techniques, something the authors call 'ethno-mining'. Similarly, Blok and Pedersen (2014) look at the intersection of 'big' and 'small' data to produce 'thick' data and include research subjects as co-producers of knowledge about themselves (see also Hsu 2014). Finally, Krieg et al. (2017) not only elaborate on the usefulness of algorithms for ethnographic fieldwork, but also show in detail how to conduct such research in an example of online reports of drug experiences.

Anthropology vs. Data Analysis

For an anthropologist, statistical and computational analysis is not the first thing that comes to mind when developing research design and methodology. Anthropologists are trained to observe phenomena in the field, talk to people, spend time with them, participate in daily activities, and immerse themselves in research

topics (Kawulich 2005; Marcus 2007). This type of information gives us detailed stories of human lives, uncovers meanings behind rituals, habits, languages, and relationships, and provides a coherent explanation of the researched phenomena. So why would anthropologists even have to include data analysis in their studies? Why and when is such an approach relevant?

Sometimes, the phenomena that anthropologists are trying to explain occur in different places at the same time and are impossible to observe simultaneously. It could be that anthropologists know little of the topic they are exploring and have yet to generate their research questions. Alternatively, the nature of the phenomenon lends itself nicely to computational analysis. For example, behaviour of many individuals is difficult to observe in real time, especially if we want to observe them at once in different locations. Sensors, on the other hand, can track behaviours of these individuals independently (Patel et al. 2012) and therefore enable a detailed comparative analysis. With a large number of measurements, researchers can also observe seasonal variations, similarity of users, and changes through time.

Data analysis also helps us define the parameters of a research field and establish what is ordinary and what extraordinary behaviour. Visualisations in particular are excellent tools for exploring and understanding frequent patterns of behaviour and outliers. When done well, visualisations harness the perceptual abilities of humans to provide visual insights into data (Fayyad et al. 2002, 4). Moreover, they provide a new perspective on a phenomenon and help generate research questions and hypotheses. Once we know how research participants behave (or communicate if we are observing textual documents or establish social ties if we are observing social networks), we can enter the field equipped with knowledge and information to verify and contextualise.

Finally, large data sets are particularly appropriate for computational analysis. While 'big data' became a popular buzzword in data science, anthropologists most likely will not be dealing with millions of data points that can be analysed only with graphics processing units (GPUs). However, even ten thousand observations are too much for a researcher to make sense of. For such data, we need software tools and visualisations, which provide an overview of the phenomenon, plot typical patterns, and enable exploring different sub-populations.

Data Ethics and Surveillance Technologies

Computational anthropology does not encompass only methodological approaches for data analysis, epistemological questions on the relationship of human being towards technology, and empirical research with computational methods, but also ethics on data storage, processing, analysis and dissemination. In a broad sense, it includes three axes of ethics, namely the ethics of data, the ethics of algorithms, and ethics of practices (Floridi and Taddeo 2016). In this contribution, we mostly touch upon the final one, the ethics of practices.

Research ethics, in particular sensitivity to the potential harm a study could elicit, is one of the core questions of anthropology, which is deeply immersed in the personal human experience. First, a solid deontological paradigm is crucial for working with not only sensitive data but any human-produced data. In this sense, we follow the principles of positivist ethics which call for human dignity, autonomy, protection, maximizing benefits and minimizing harm, respect, and justice (Markham et al. 2012; Halford 2017). In other words, anthropologists should act in the best interest of the research participants and avoid or minimize negative effects the study could have on the people.

Secondly, anthropologists should be mindful of the potential subjectivity of their interpretation of the data. Every data set, whether quantitative or qualitative, elicits interpretation that inevitably stems from our own world-view. To keep the bias to a minimum, the suggested circular mixed methods approach, proposed in this article, as well as most others approaches with origins in anthropology strive for continual reinterpretation of the results within the actual social context of research participants. Each ethnographic layer explains the results from the point of view of data producers and thus minimizes the chance of bias and misinterpretation.

Sensors and wearable devices inevitably invoke questions of surveillance and privacy. Here, we propose a distinction between surveillance and monitoring. Surveillance implies guiding actions of surveilled subjects, while monitoring proposes a more passive stance of observing behaviour (see Marx 2002; Nolan 2018). The present study was not designed to guide behaviour but to observe and understand, hence being more monitoring than surveillance focused. And even if we consider it surveillance-like, Marx proposes "a broad comparative measure of surveillance slack which considers the extent to which a technology is applied, rather than the absolute amount of surveillance" (Marx 2002), meaning that the extent to which surveillance is harmful is the power it holds for the user. The case of sensor data of a smart building that monitors only neutral human behaviour, falls to the soft side of power, which, in the opinion of the authors, deserves some surveillance slack. Nevertheless, we strived to uphold high ethical standards for handling the data and disseminating the results, mostly by employing "ongoing consensual decision-making" (Ramos 1989) by informing participants of the purpose of the research, which data are being collected and how the findings are going to be presented.

Circular Mixed Methods

Circularity in machine learning and data mining is not a novel idea. Data science methodology already includes ideas about circular phases of data mining (CRISP-DM, Shearer 2000), where phases are interdependent and by reiterating through them the analyst clarifies existing and generates new business questions. However, not much has been written yet alone put to practice in terms of interdisciplinary circularity and the intertwining of methods from different scientific fields (for some pioneering efforts see above).

For the purpose of our study, we designed a novel methodological approach, named it *circular mixed methods*, and employed it to analyse the workspace behaviour and practices of employees. This approach aims to observe the phenomenon from several different perspectives. Nominally, we have split these perspectives into several research stages, where we use a single method, but in reality these methods are used interchangeably and in accordance with each particular situation. For the sake of clarity, however, we will refer to the four stages of research.

The first stage involved gathering historical longitudinal data from the building's sensors. We used unsupervised data mining and exploratory data analysis to uncover behavioural patterns, identify interesting individuals (outliers) and form several hypotheses about the use of spaces and energy consumption in the building.

The second stage involved in-depth ethnographic research, where we used interviews, questionnaires, focus groups, and, most importantly, participant observation from a three-year period of working in the mentioned building. This part helped us clarify the context of particular behaviours, identified the values, motivations and deterring factors of each research participant, and confirmed or rejected hypotheses.

In the third stage, which is currently in process, we use text mining methods on interview transcripts to find common topics, observe sentiment towards particular issues, and determine which individuals have similar opinions on certain topics. This is still a work-in-progress and the results might even be negative – since this is the first application of text mining on ethnographic interview transcripts, we still need to estimate the viability of such an approach for anthropology. The size of the corpus in anthropology is normally small and it is entirely possible there is no added value to text mining such data. Finally, we conclude with another round of ethnography, conducting the second round of interviews (normally a year after the first one) and verifying the results from the third stage.

As mentioned before, the distinction between each particular stage is not always strict. Circular mixed methods aim to provide the researcher with the freedom she needs to address complex phenomena from several different perspectives. There can be several alternating stages, where each stage contributes another interpretative layer to the previously established facts.

There are plenty of benefits of this approach. Circular mixed methods are particularly appropriate for uncovering intricate longitudinal patterns, which are incredibly challenging for the researcher to observe at such granularity. Moreover, this method can be used for observing diachronic phenomena, where the data comes from several locations at once, hence overcoming the physical limitations of a single researcher. Finally, researchers can effectively and rapidly analyse large data collections with computational means. Some such collections are interesting for anthropologists as well, namely social media, archival data, wearables, sensors or audio-visual recordings. Visualisations, one of the product of data mining, substantiate the findings and enable researchers to uncover relations, patterns and outliers in the data. Data analysis can thus help generate hypotheses and questions for the research. This cuts down the time

required to get familiar with the field. A researcher can come into the field equipped with potentially interesting hypotheses and test them almost immediately.

Looking at the data alone, however, we would be unable to determine what any of those patterns and outliers mean. To truly understand them, we need to immerse ourselves in the field, ask questions and observe how people behave and create their habits and practices. While quantitative analysis provides us with clues, qualitative approaches, such as ethnography and fieldwork, explain those clues and substantiate the superficial knowledge of the field acquired in the first and third research phase. Metaphorically speaking, data analysis is great for scratching the surface and ethnography excels at digging deeper. By combining the two approaches, however, we can interpret the data in a rich and meaningful way, as we will show with the case study presented in the article.

Data Preprocessing

In our study, we have observed sensor measurements from a faculty building which is considered to be a state-of-the-art smart building in Slovenia. Each room in the building is equipped with a temperature sensor and sensors on windows that track when they are open or closed. Doors have electronic key locks that track when the room is occupied. There were altogether 11 sensor measurements, with additional 8 measurements coming from the weather station located on the building's rooftop. In-room sensor reports the room temperature, set temperature, ventilation speed, daily regime, and so on, while the weather station reports the external temperature, light, rainfall, etc.

One of the most important measures is the daily regime, which has four values, each representing a state of the overall room setting. When a person is present in the room, the regime is comfort (value = 0) and when a window is open, the regime is off (value = 4). If the room is vacant, the regime goes to night (value = 1) or standby (value = 3).[2] These measurements come from electronic locks on the doors, which record when the room is occupied, and the magnets in windows, which record when the window is open.

We retrieved 55,456 recordings for 14 rooms of different types, namely 5 laboratories, 6 cabinets, and 3 administration rooms. Measurements are recorded bi-hourly and stored in SCADA, a software that allows controlling processes locally or at remote locations, monitoring and processing real-time data, interacting with devices, and recording events into a log file.

We decided to observe the year 2016 and later compare it to 2017. The results in the paper refer only to 2016. The rooms are anonymised to ensure data privacy and results for two of the rooms are not reported at the request of their occupants.

2 Standby is activated on workdays as a transitory setting between night and comfort regime.

Table 1: Original data.

Date	Room temperature	Daily regime	Room
2016-01-01 02:10:00	20.94266	1	C
2016-01-01 02:10:00	21.65854	1	B
2016-01-01 02:10:00	20.63234	1	K
2016-01-01 02:10:00	22.41270	1	D
2016-01-01 02:10:00	20.25890	1	M
2016-01-01 02:10:00	21.45220	3	C

Source: Author data.

We performed extensive data cleaning and preprocessing and removed data points with missing values (Table 1). The daily regime was considered the most important variable since it reports a presence in the room or the opening of windows. Concurrently, we retained only the variables reporting the daily regime, since this feature registered human behaviour the best, and room temperature. We also generated additional features, such as the day of the week and room type (cabinet, laboratory, and administration).

Table 2: Data transformed into a behaviour vector. 1 denotes occupancy of the room, meaning daily regime value was either 0 (comfort) or 4 (window open).

Date	0 am	1 am	2 am	3 am	…	Room	Day	Type
2016-01-01	0	1	1	0	…	C	Fri	laboratory
2016-01-01	0	0	0	0	…	B	Fri	laboratory
2016-01-01	0	0	1	1	…	K	Fri	administration
2016-01-01	0	0	0	1	…	D	Fri	cabinet
2016-01-01	0	1	1	1	…	M	Fri	administration
2016-01-02	0	0	1	1	…	C	Sat	laboratory

Source: Author data.

In the second part of the analysis, we created a transformed data set where we merged daily readings for a room into one 'daily behaviour' vector (Table 2). In the new data set, each room has a daily recording, where the new features are values of the daily regime at each hour. Since sensors only record the state every two hours, we filled missing values with the previously observed state. For example, if the original vector was {0, ?, 0, ?, 1, ?, 1}, we imputed missing values to get {0, 0, 0, 0, 1, 1, 1}. As we were interested only in the presence in the room, we put 0 where daily regime was 1 (night) or 3 (standby) and 1 where it was 0 (comfort) or 4 (window open), discarding the information on specific temperature regimes. This gave us the final daily behaviour vector which we could compare in time and between rooms.

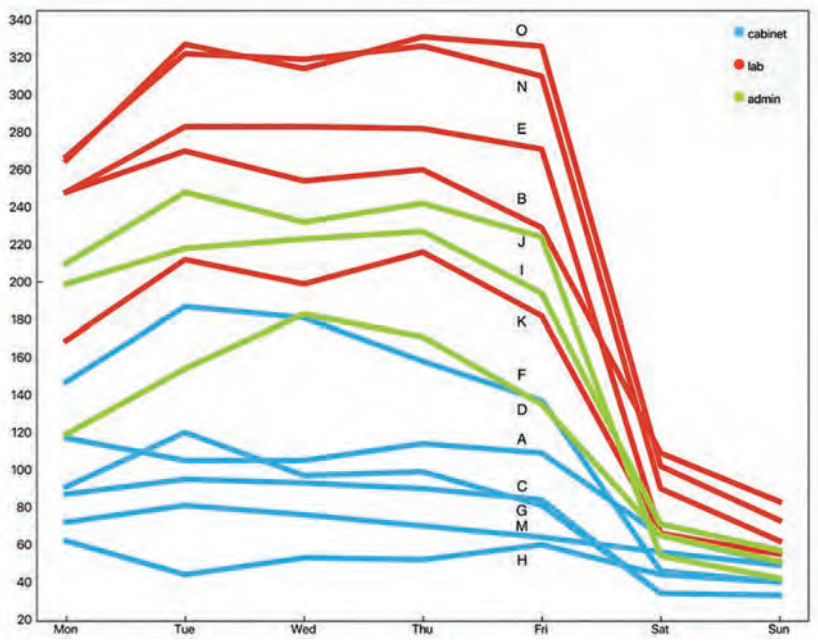
To sum up, we were working with two data sets, the first reporting the presence of people in the room for a given time (11 features from room sensors) and the second

one showing the behaviour of the room throughout the day (24 features on room occupancy at each hour).

Results

First, we wanted to see how rooms differ by room occupancy alone. We hypothesised there will be a significant difference in occupancy between laboratories and cabinets since the presence of more people in a space extends the occupancy hours (no complete overlap of working time). We took the first data set with bi-hourly recordings and removed readings where the daily regime was either 1 (night) or 3 (standby) because these readings indicate the room was not occupied. Afterwards, we computed the contingency matrix of room occupancy by the day of the week, which shows how many times per year a room was occupied on a certain day. We visualised the result in a line plot by the type of the room (Figure 1). We can notice that laboratories have a higher presence on Saturday and Sunday than the other rooms.

Figure 1: Occupancy of the rooms for each day of the week



Source: Author data displayed in the software Orange.

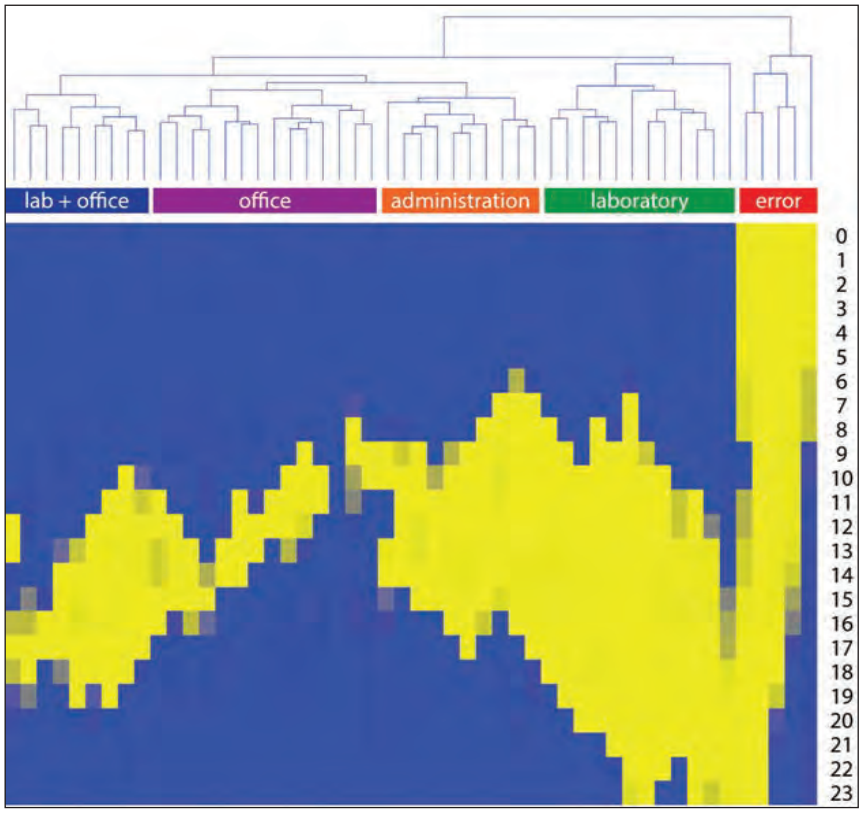
Moreover, N and O are the top two rooms by occupancy. We know that these two rooms belong to a single laboratory and are separated with a permanently open door. These two rooms are occupied by the largest number of people and since the

employees of the faculty have a somewhat flexible working time, the dispersion of working time is expectedly the highest in rooms with the most occupants (smallest overlap in working time among employees). N and O are also among the few rooms where occupancy goes up towards the end of the week.

F and B are also laboratories, both displaying similarly high presence across the week. On the bottom of the plot there are cabinets, namely G, K, F. Unsurprisingly, cabinets display lower occupancy rates than laboratories, since cabinets are used by a single person and hence no overlap is possible. They are also functional rooms, used predominantly for meetings, office hours, and other intermittent work of professors.

With the second room occupancy data set, we made an analysis of behavioural patterns by the time of the day. We observed occupancy by room type in a heat map where 1 (yellow) means presence and 0 (blue) absence. Visualisation in Figure 2 is simplified by merging similar rows with k-means (k = 50) and clustering by similarity (Euclidean distance, average linkage, and optimal leaf ordering). Such simplification joins identical or highly similar patterns into one row and rearranges them so that similar rows are put closer together.

Figure 2: Occupancy of the rooms for each day of the week

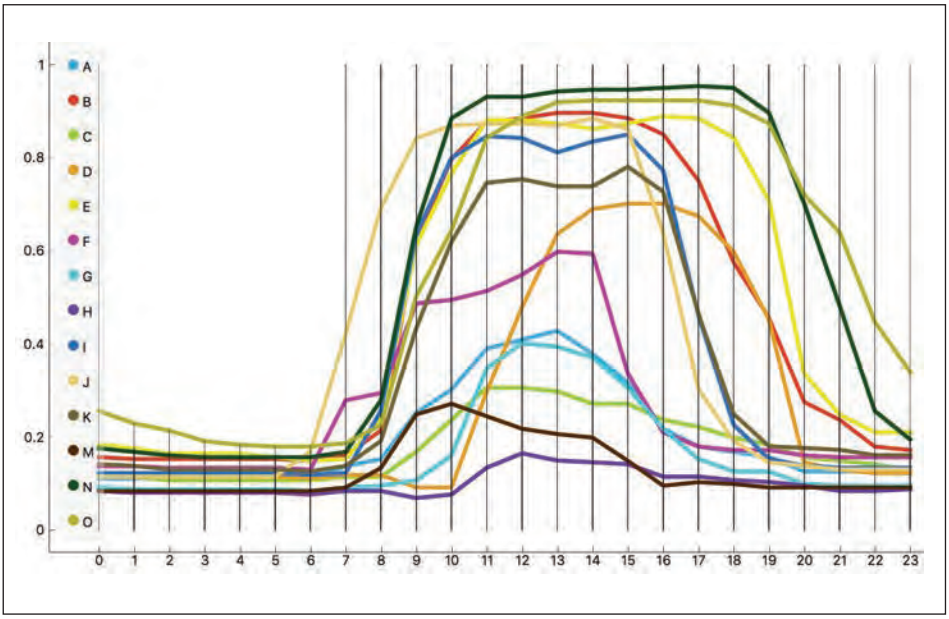


Source: Author data displayed in the software Orange.

Clustering revealed that occupancy sequence highly depends on the room type. There were some error data, where sensors recorded presence at unusual hours (for example during the night consistently across all rooms). But despite some noise in our data, we can distinguish between typical laboratory, administration and cabinet behaviour, since our error data constitute a separate cluster (Dave 1991). Cabinets again show the lowest occupancy with presence recorded sporadically across the day. Normally, university lecturers spend a large portion of their time in lecture rooms and in their respective laboratories. This is why occupancy of cabinets is so erratic and does not display a consistent pattern. Laboratory occupants, on the other hand, usually come late and stay late, while administration staff work regularly from 7:00 a.m. to 4:00 p.m. They both display fairly consistent behaviour.

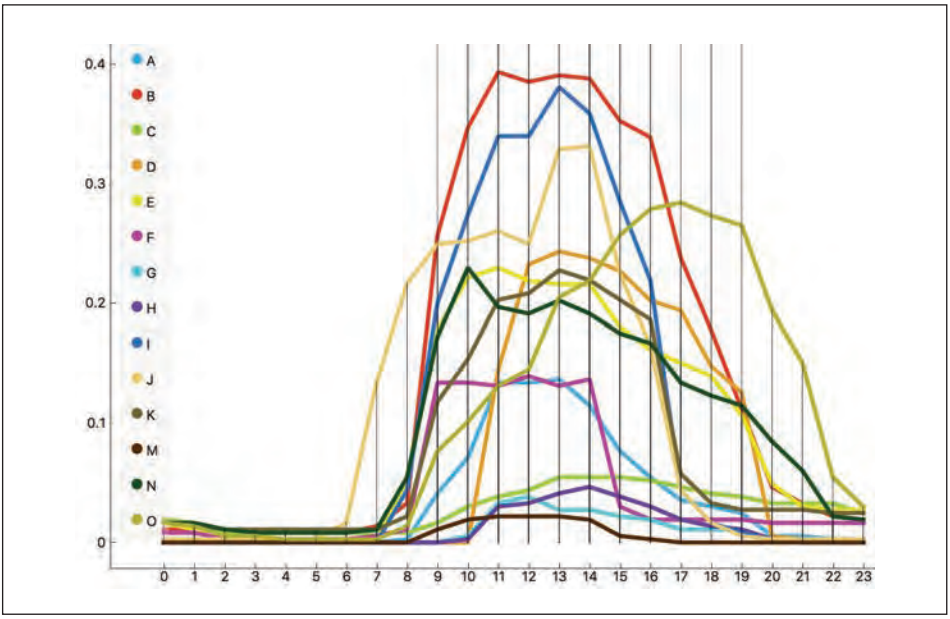
We visualised the same data set in a line plot, which shows the frequency of attributes on a line. In this way, we can better observe the differences between individual rooms at each time of the day and where specific peaks (high frequencies) happen. Figure 3 displays the occupancy ratio at a specific time of day, while Figure 4 shows the ratio of window opening.[3] Several interesting observations emerge. In both cases, room O is skewed to the right, meaning its occupants work at late hours and open windows while working. Conversely, room J is skewed to the right, indicating its occupants start work earlier than most. There is also a distinct peak in window opening at around lunch time.

Figure 3: Room occupancy by the time of day



3 1 would mean the room was always occupied and 0 that the room was never occupied at a specific time of the day.

Figure 4: Window opening frequency by the time of the day



In most rooms, people are opening windows from late morning to early afternoon. Again, not surprising, considering this is their peak working time. This is a great indicator for an ethnographer if he or she wants to observe windows interaction (who does it, is there a consensus on whether or not it should be opened, does this happen more frequently after lunch …). Looking at the data, the best time for observing the specified behaviour is between 10:00 a.m. and 1:00 p.m. Accordingly, data analysis can also serve as a guide for ethnographic fieldwork.

Ethnography Comes In

Data analysis revealed some interesting patterns in the use of working spaces:

- laboratories work more on the weekends,
- rooms N and O work late,
- room J starts the day early and opens the windows at lunchtime, and
- in rooms H, N and O the occupancy goes up towards the end of the week.

How can we explain this? While the data gave us clues, the answers lie with the people. Substantiating analytical findings with fieldwork ethnography is crucial for understanding the data. We conducted semi-structured interviews with the rooms' occupants to discover what those patterns mean and why a certain behaviour occurs.

Laboratories have a higher weekend occupancy since they offer a quiet place to work for PhD students who are either catching deadlines for publishing papers or using their 'off time' for some in-depth research. Room B, in particular, seems to like working at weekends and we were able to identify an individual who often comes to work on Saturdays. In the interview, he[4] told us this was the time when he finally managed to do some actual work: *"Effectively, if you look at the duration of my focus, it is much longer during the weekends. In my opinion, I do a day and a half worth of work during the weekend compared to the weekday."*

Rooms N and O are quite similar in terms of presence although room N displays a tendency to work the latest. By observing the inhabitants in this room and talking to them, we identified an individual who preferred to work in the late afternoon and evening. Since, as mentioned above, working time is flexible at the studied faculty, he adjusted his working hours to suit his preferences. He also prefers fresh air to artificial ventilation and opens the windows whenever possible. This accounts for the skew to the right for room O in Figure 4. "*I like fresh air,"* he told us. *"The air outside is always better than the air inside. I opened my window at every chance, even during the winter."*

The increased productivity in rooms N, O, and H towards the end of the week is explained by the fact that Fridays are working sprints for the occupants of these three rooms. The case of room H is particularly interesting. This is the room with the overall lowest occupancy, yet the room is most frequented on Fridays, unlike in most other rooms, where the occupancy decreases towards the end of the week. Room H is the cabinet of a professor who runs laboratories N and O. He is also a part of the Friday development sprints, hence the peak. Yet he is very sociable and prefers to work in the laboratory with colleagues, rather than alone in the cabinet, as was evident from our observation and discussion with him. This also explains the overall low and erratic occupancy of his room during the rest of the week.

The skewed peak for room J in Figure 3 is also interesting. The occupant of this room admitted he prefers coming to work earlier to make the most of the day. He stressed several times that daylight is important to him and by shifting working time to earlier hours, he was able to leave early and use the rest of the day for himself. He also said he was the most productive in early mornings since these were the quietest parts of the day. In his words: *"[I like coming early] because I have more of the day left in my private life. It is also quiet in the morning and I can do more work."*

Personal preferences evidently affected the discovered patterns of workday behaviour. In summary, people working in the researched building adjust their working hours and their environment to suit their personal needs, values, and lifestyle. Designing a single solution for such a diverse group not only invokes dissatisfaction among occupants, but leads to lower productivity, higher stress, improvised DIY solutions, and ultimately to higher energy consumption and worse workspace health.

4 For concealing the actual identity of the people participating in the study, the pronoun he is used to denote both males and females.

Conclusion

In this paper, we have shown how anthropological (qualitative) research methods can be enriched, upgraded, and substantiated by data analysis. While the findings are still preliminary and based on a limited sample, they nevertheless pinpoint aspects of data analysis that benefit from ethnographic insight and vice versa.

With the increasing availability of data, especially from sensors, wearable devices, and social media, anthropologists can use computational methods and data analysis to uncover common patterns of human behaviour and pinpoint interesting outliers. Quantitative methods have proven useful when dealing with large data sets. In such cases, an analysis without digital tools is virtually impossible, while visualisations offer new insight into the problem and help present the data concisely. In addition, quantitative approaches also increase the reproducibility of research.

However, patterns emerging from such analysis can hardly ever be explained with data alone. We argue that data analysis can generate new hypotheses and research questions (Krieg et al. 2017) and provide a general overview of the topic. Conversely, ethnography substantiates analytical findings with the context and story behind the data. Going back and forth, from quantitative to qualitative methods and approaches, enables researchers to establish a research problem as suggested by the data, gauge new perspectives on the known problems, and account for outliers and patterns in the data. Circular research design enhances the quality of information, which does not have to derive solely from a quantitative or qualitative approach. By combining the two, we are using a research loop that ensures both sets of data get an additional perspective – quantitative data are verified with ethnography in the field, while ethnographic data become supported with statistically relevant patterns analysed by computational tools.

Such methods are already, to a certain extent, employed in digital anthropology (Drazin 2012), but they are gaining more prominence in mainstream anthropology as well (Krieg et al. 2017). By establishing a solid methodological framework for quantitative analyses in relation to qualitative ones, we do not only strengthen the subfield of computational anthropology, but also provide new perspectives and research ventures to anthropology and emphasise its relevance for understanding lifestyles, habits, and practices in data-driven societies.

Sources and Literature

Literature:

- Anderson, Ken, Dawn Nafus, Tye Rattenbury, and Ryan Aipperspach. 2009. "Numbers Have Qualities Too: Experiences with Ethno-mining." *Ethnographic Praxis in Industry Conference Proceedings* 2009 (1): 123–40.
- Bell, Genevieve. 2006. "Satu keluarga, satu komputer (One Home, One Computer): Cultural Accounts of ICTs in South and Southeast Asia." *Design Issues* 22 (2): 35–55.

- Blok, Anders, and Morten Axel Pedersen. 2014. "Complementary Social Science? Quali-quantitative experiments in a Big Data World." *Big Data & Society* 1 (2): 1–6.
- Boellstorff, Tom. 2015. *Coming of Age in Second Life: An Anthropologist Explores the Virtually Human*. Princeton: Princeton University Press.
- Bonilla, Yarimar, and Jonathan Rosa. 2015. "#Ferguson: Digital Protest, Hashtag Ethnography, and the Racial Politics of Social Media in the United States." *American Ethnologist* 42 (1): 4–17.
- boyd, danah, and Kate Crawford. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society* 15 (5): 662–79.
- Conte, Rosaria, Nigel Gilbert, Giulia Bonelli, Claudio Cioffi-Revilla et al. 2012. "Manifesto of Computational Social Science." *The European Physical Journal Special Topics* 214 (1): 325-346
- Creswell, John W., and Vicki L. Plano Clark. 2007. *Designing and Conducting Mixed Methods Research*. Thousand Oaks: Sage publications.
- Dave, Rajesh N. 1991. "Characterization and Detection of Noise in Clustering." *Pattern Recognition Letters* 12 (11): 657–64.
- Dobbert, Marion Lundy, Dennis P. McGuire, James J. Pearson, and Kenneth Clarkson Taylor. 1984. "An Application of Dimensional Analysis in Cultural Anthropology." *American Anthropologist* 86 (4): 854–84.
- Drazin, Adam. 2012. "Design Anthropology: Working on, with and for Digital Technologies." In *Digital Anthropology*, edited by Heather A. Horst and Daniel Miller, 245–65. London and New Yourk: Berg.
- Fayyad, Usama M., Andreas Wierse, and Georges G. Grinstein, eds. 2002. *Information Visualization in Data Mining and Knowledge Discovery*. San Francisco: Morgan Kaufmann.
- Floridi, Luciano and Taddeo, Mariarosaria. 2016. "What is Data Ethics?" *Philosophical Transactions A*: 374: 1–5.
- Gershenfeld, Neil, and J. P. Vasseur. 2014. "As Objects go Online: The Promise (and Pitfalls) of the Internet of Things." *Foreign Affairs* 93: 60.
- Halford, Susan. 2017. "The Ethical Disruptions of Social Media Data: Tales from the Field." In *The Ethics of Online Research*, 13–25. Bingley: Emerald Publishing Limited.
- Hockey, Susan. 2004. "The History of Humanities Computing." *A Companion to Digital Humanities*, 3–19.
- Hsu, Wendy F. 2014. "Digital Ethnography Toward Augmented Empiricism: A New Methodological Framework." *Journal of Digital Humanities* 3 (1): 1–19.
- Jänicke, Stefan, Greta Franzini, Muhammad Faisal Cheema, and Gerik Scheuermann. 2015. "On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges." In *Eurographics Conference on Visualization (EuroVis)-STARs*. The Eurographics Association.
- Johnson, R. Burke, Anthony J. Onwuegbuzie, and Lisa A. Turner. 2007. "Toward a Definition of Mixed Methods Research." *Journal of Mixed Methods Research* 1 (2): 112–33.
- Kawulich, Barbara. 2005. "Participant Observation as a Data Collection Method." In *Qualitative Sozialforschung / Forum: Qualitative Social Research* 6 (2).
- Klein, Lauren F., and Matthew K. Gold. 2016. "Digital Humanities: The Expanded Field." *Debates in the Digital Humanities*.
- Krieg, Lisa Jenny, Moritz Berning, and Anita Hardon. 2017. "Anthropology with Algorithms? An Exploration of Online Drug Knowledge Using Digital Methods." *Issues* 5 (2).
- Kuzara, Richard S., George R. Mead, and Keith A. Dixon. 1966. "Seriation of Anthropological Data: A Computer Program for Matrix-ordering." *American Anthropologist* 68 (6): 1442–55.
- Lum, Kristian, and William Isaac. 2016. "To Predict and Serve?" *Significance* 13: 14–19.
- Malinowski, Bronislaw. 2002 [1922]. *Argonauts of the Western Pacific: An Account of Native Enterprise and Adventure in the Archipelagoes of Melanesian New Guinea*. London: Routledge.
- Marcus, George E. 2007. "Ethnography Two Decades after Writing Culture: From the Experimental to the Baroque." *Anthropological Quarterly* 80 (4): 1127–45.

- Markham, Annette, Elizabeth Buchanan, and AoIR Ethics Working Committee. 2012. *Ethical Decision-making and Internet Research: Version 2.0.* Association of Internet Researchers.
- Marx, Gary T. 2002. "What's New about the 'New Surveillance'? Classifying for Change and Continuity." *Knowledge, Technology & Policy* 17 (1): 18–37.
- Miller, Daniel, Elisabetta Costa, Nell Haynes, Tom McDonald, Razvan Nicolescu, Jolynna Sinanan, Juliano Spyer, Shriram Venkatraman, and Xinyuan Wang. 2016. *How the World Changed Social Media*. London: UCL press.
- Mitchell, J. Clyde. 1974. "Social Networks." *Annual Review of Anthropology* 3: 279–99.
- Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. "The Ethics of Algorithms: Mapping the Debate." *Big Data & Society* 3 (2): 1–21.
- Nardi, Bonnie. 2010. *My Life as a Night Elf Priest: An Anthropological account of World of Warcraft*. University of Michigan Press.
- Nolan, Cathy. 2018. "Data Surveillance, Monitoring, and Spying: Personal Privacy in a Data-Gathering World." Data Topics. Published May 2, 2018. https://www.dataversity.net/data-surveillance-monitoring-spying-personal-privacy-data-gathering-world/.
- Patel, Shyamal, Hyung Park, Paolo Bonato, Leighton Chan, and Mary Rodgers. 2012. "A Review of Wearable Sensors and Systems with Application in Rehabilitation." *Journal of Neuroengineering and Rehabilitation* 9 (1): 21.
- Pink, Sarah, Shanti Sumartojo, Deborah Lupton, and Christine Heyes La Bond. 2017. "Mundane Data: The Routines, Contingencies and Accomplishments of Digital living." *Big Data & Society* 4 (1): 1–12.
- Podolefsky, Aaron, and Christopher McCarty. 1983. "Topical Sorting: A Technique for Computer Assisted Qualitative Data Analysis." *American Anthropologist* 85 (4): 886–90.
- Pretnar, Ajda, and Marko Robnik-Šikonja. 2019. "Analiza slik in besedil s pristopi umetne inteligence." *Glasnik SED* 59 (1): 49–57.
- Ramos, Mary Carol. 1989. "Some Ethical Implications of Qualitative Research." *Research in Nursing & Health* 12 (1): 57–63.
- Shearer, Colin. 2000. "The CRISP-DM Model: the New Blueprint for Data Mining." *Data Warehousing* 5: 13–22.
- Skeem, Jennifer L., and Christopher Lowenkamp. 2016. "Risk, Race, & Recidivism: Predictive Bias and Disparate Impact." *Criminology: An Interdisciplinary Journal* 54 (4): 680–712.
- Svensson, Patrik. 2010. "The Landscape of Digital Humanities." *Digital Humanities*. http://digitalhumanities.org/dhq/vol/4/1/000080/000080.html.
- Teddlie, Charles, and Abbas Tashakkori. 2009. *Foundations of Mixed Methods Research: Integrating Quantitative and Qualitative Approaches in the Social and Behavioral Sciences*. Los Angeles: Sage.
- White, Douglas R., and Gregory F. Truex. 1988. "Anthropology and Computing: The Challenges of the 1990s." *Social Science Computer Review* 6 (4): 481–97.

Oral Sources:

- Personal archive.

Ajda Pretnar, Dan Podjed

DATA MINING WORKSPACE SENSORS: A NEW APPROACH TO ANTHROPOLOGY

SUMMARY

With an increasing availability of data coming from social networks and wearable devices among other sources, anthropologists can easier than ever dive into data analysis and study humans and their societies, subcultures and cultures quantitatively as well as qualitatively. In this contribution we extend the interdisciplinarity of anthropology by employing circular mixed methods that combine qualitative (ethnographic) approaches with quantitative approaches from data mining and machine learning.

The research, which is the basis for this contribution, began in October 2017 and includes 14 workspaces of one of University of Ljubljana's buildings. For the purpose of our study, we designed a novel methodological approach and named it *circular mixed methods*. We employed it to analyse workspace behaviours and practices of employees and to develop sustainable solutions for encouraging a healthy lifestyle.

As we explain in the contribution, circular mixed methods are appropriate for uncovering detailed longitudinal patterns, which are impossible to detect manually. The suggested approach is effective for analysing diachronic phenomena, where we retrieve the data from several locations at once, thus overcoming the physical limitations of individual researchers. Finally, using this methodology, researchers can effectively and rapidly analyse large data collections. Some such collections are interesting for anthropologists as well, namely social media, archival data, wearables, sensors or audio-visual recordings.

While quantitative analysis helps us generate hypotheses and uncover patterns in the data, qualitative approaches, such as ethnography and fieldwork, explain those patterns and substantiate data with rich details. Combining the two approaches, we can interpret the data in a contextually rich and anthropologically relevant way.

Ajda Pretnar, Dan Podjed

PODATKOVNO RUDARJENJE SENZORJEV V DELOVNEM OKOLJU: NOV PRISTOP K ANTROPOLOGIJI

POVZETEK

Z vse večjo množico podatkov, pridobljenih, med drugim, z družbenih omrežij in pametnih naprav, lahko antropologi lažje kot kadarkoli prej pri raziskovanju uporabijo podatkovno analitiko in preučujejo ljudi in njihove navade ter kulture in podkulture, in to tako s kvantitativnega kot kvalitativnega vidika. V prispevku razširimo idejo interdisciplinarnosti v antropologiji z uporabo krožnih mešanih metod, ki povezujejo kvalitativne (etnografske) pristope s kvantitativnimi pristopi rudarjenja podatkov in strojnega učenja.

Raziskava, ki je podlaga pričujočega prispevka, se je začela oktobra 2017 in vključuje 14 delovnih prostorov ene od stavb Univerze v Ljubljani. Za vzpostavljanje novih pogledov na navade ljudi v stavbi in snovanje potencialno relevantnih hipotez smo uporabili *krožne mešane metode*, s katerimi smo analizirali vedenje in prakse ter na podlagi le-teh razvili celostne rešitve za spodbujanje zdravega načina življenja in izboljšanje počutja na delovnem mestu.

Kot pojasni prispevek, so krožne mešane metode najprimernejše za odkrivanje podrobnih in dolgotrajnih vzorcev, ki jih raziskovalec ne more sam opazovati in zaznati. Pristop je učinkovit tudi za opazovanje sočasnih dogodkov, kjer podatke hkrati pridobimo z več lokacij, s čimer presežemo raziskovalčeve fizične omejitve. Poleg tega je metodologija uporabna za učinkovito in hitro analizo velikih podatkovnih zbirk, med katerimi so za antropologe posebej zanimivi podatki z družbenih omrežij, pametnih naprav in senzorjev, avdio-vizualno gradivo ter digitalizirani arhivski viri.

Medtem ko kvantitativna analiza omogoča postavitev hipotez in odkrivanje vzorcev v podatkih, jih kvalitativne metode, zlasti etnografija in opazovanje na terenu, razložijo in obogatijo s podrobnostmi. Kombinacija obeh pristopov zagotavlja, da podatke interpretiramo na vsebinsko bogat in antropološko relevanten način.

1.01 UDC: 070.18:329.052 (497.5)"1945/1990"

Tadej Škvorc,[*] Simon Krek,[**] Senja Pollak,[***]
Špela Arhar Holdt,[****] Marko Robnik-Šikonja[*****]

Predicting Slovene Text Complexity Using Readability Measures

IZVLEČEK

NAPOVEDOVANJE KOMPLEKSNOSTI SLOVENSKIH BESEDIL
Z UPORABO MER BERLJIVOSTI

Večina obstoječih formul za merjenje berljivosti je zasnovana za besedila v angleškem jeziku, na katerih je tudi ocenjena njihova kakovost. V našem članku predstavimo prilagoditev izbranih mer za slovenščino. Uspešnost desetih znanih formul ter osmih dodatnih kriterijev berljivosti ocenimo na petih skupinah besedil: otroških revijah, splošnih revijah, časopisih, tehničnih revijah in zapisnikih sej državnega zbora. Te skupine besedil imajo različne ciljne publike, zaradi česar predpostavimo, da uporabljajo različne stile pisanja, ki bi jih formule in kriteriji berljivosti morali zaznati. V analizi pokažemo, katere formule in kriteriji berljivosti delujejo dobro in s katerimi razlik med skupinami nismo mogli zaznati.

Ključne besede: berljivost, obdelava naravnega jezika, analiza besedil

[*] **University of Ljubljana, Faculty of Computer and Information Science, Večna Pot 113, SI-1000 Ljubljana, Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, tadej.skvorc@fri.uni-lj.si**

[**] **Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, University of Ljubljana, Faculty of Arts, Aškerčeva 2, SI-1000 Ljubljana, simon.krek@guest.arnes.si**

[***] **Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, senja.pollak@ijs.si**

[****] **University of Ljubljana, Faculty of Arts, Aškerčeva 2, SI-1000 Ljubljana, University of Ljubljana, Faculty of Computer and Information Science, Večna Pot 113, SI-1000 Ljubljana, spela.arharholdt@ff.uni-lj.si**

[*****] **University of Ljubljana, Faculty of Computer and Information Science, Večna Pot 113, SI-1000 Ljubljana, marko.robnik@fri.uni-lj.si**

ABSTRACT

The majority of existing readability measures are designed for English texts. We aim to adapt and test the readability measures on Slovene. We test ten well-known readability formulas and eight additional readability criteria on five types of texts: children's magazines, general magazines, daily newspapers, technical magazines, and transcriptions of national assembly sessions. As these groups of texts target different audiences, we assume that the differences in writing styles should be reflected in their readability scores. Our analysis shows which readability measures perform well on this task and which fail to distinguish between the groups.

Keywords: readability, natural language processing, text analysis

Introduction

In English, the problem of determining text readability (i.e. how easy a text is to understand) has long been a topic of research, with its origins in the 19th century (Sherman 1893). Since then, many different methods and readability measures have been developed, often with the goal of determining whether a text is too difficult for its target age group. Even though the question of readability is complex from a linguistic standpoint, a large majority of existing measures are based on simple heuristics. There has been little research on readability of languages other than English, therefore we aim to apply these measures to Slovene and evaluate how well they perform.

There are several factors that might cause these measures to perform poorly on non-English languages, such as:

- Many measures are fine-tuned to correspond to the grade levels of the United States education system. It is likely a different fine-tuning would be needed for other languages, as a.) their education system is different from the US system, and b.) the differences in readability between grade levels are likely to be different between languages, meaning that each language would require specifically tuned parameters.
- Some measures utilize a list of common English words and their results depend on the definition of this list. For Slovene, there currently does not exist a publicly available list of common words, so it is not known how such measures would perform.
- The existing readability measures do not use the morphological information to determine difficult words but rely on syllable and character counts, or a list of difficult words. As Slovene is morphologically much more complex than English, words with complex morphology are harder to understand than those with simple morphology, even if they have the same number of characters or syllables.

We analyze the commonly used readability measures (as well as some novel measures) on Slovene texts and propose a word list needed to implement the word-list-based measures. We calculate statistical distributions of scores for each readability measure across subcorpora and assess the ability of measures to distinguish between different subcorpora using a variety of statistical tests. We show that machine learning classification models, using a combination of readability measures, can predict the subcorpus a given text belongs to.

The paper extends the short version of the paper presented in Škvorc et al. (2018) and is structured as follows. We first present the related work on readability measures and describe the readability measures used in our analysis. The methodology of the analysis is presented next, followed by the results split into three sections. The last section concludes the paper and presents ideas for further work.

Related Work

For English, there exists a variety of works focused on determining readability by using readability formulas. Those formulas rely on different features of the text such as the average sentence length, percentage of difficult words, and the average number of characters per word. Examples of such measures include the Coleman-Liau index (Coleman and Liau 1975), LIX (Björnsson 1968), and the automated readability index (ARI) (Senter and Smith 1967). Some formulas, like the Flesch-Kincaid grade level (Kincaid et al. 1975) and SMOG (Mc Laughlin 1969) use the number of syllables per word to determine if a word is difficult. Additionally, some measures (e.g., the Spache readability formula (Spache 1953) and Dale-Chall readability formula (Dale and Chall 1948) rely on a pre-constructed list of difficult words.

Aside from the readability formulas, there exists a variety of other approaches that can be used to determine readability (Bailin and Grafstein 2016). For example, various machine-learning approaches can be used to obtain better results than readability formulas, such as the approach presented in Francois and Miltsakaki (2012), which outperforms readability formulas on French text.

There is little work attempting to apply these measures to Slovene texts. Most work dealing with the readability of Slovene text is focused on manual methods. For example, Justin (2009) analyzes Slovene textbooks from a variety of angles, including readability. On the other hand, works that focus on automatic readability measures are rare. Zwitter Vitez (2014) uses a variety of readability measures for author recognition in Slovene text, but we found no works that used them to determine readability.

In addition to Slovene, some related works evaluate readability measures on other languages. Debowski et al. (2015) evaluate readability formulas on Polish text and show that they obtain better results by using a more complex, machine-learning-based approach.

Readability Measures

In our analysis, we used two groups of readability measures:

- **Existing readability formulas for English:** we focused mainly on popular methods that have been shown to achieve good results on English texts. These measures mostly rely on easy-to-obtain features such as a number of difficult words, sentence length, and word length.
- **Natural-language-processing-based readability criteria:** we used additional criteria that are not present in the existing readability formulas but can be obtained from tools for automatic language processing, such as the percentage of verbs, number of unique words, and morphological difficulty of words. In the existing English formulas, such criteria are not used but they might contain useful information for determining the readability of Slovene texts.

In the following two subsections we present the established readability measures for grading English text and our proposed additional criteria.

Existing Readability Formulas

There exists a variety of ways to measure the readability of texts written in English. For our analysis, we used 10 readability formulas given below. The entities used in the expressions correspond to the number of occurrences of a given entity, e.g., word corresponds to the number of words in a measured text.

- **Gunning fog index** (Gunning 1952) is calculated as:

$$\mathrm{GFI} = 0.4\frac{\text{words}}{\text{sentences}} + 100\frac{\text{complex words}}{\text{words}},$$

where a word is considered complex if it contains three or more syllables. As there exists no established automatic method for counting syllables of Slovene words, we used a rule-based approach designed for English. The resulting score is calibrated to the grade level of the USA education system.

- **Flesch reading ease** (Kincaid et al. 1975) is calculated as:

$$\mathrm{FRE} = 206.835 - 1.015\frac{\text{words}}{\text{sentences}} - 84.6\frac{\text{syllables}}{\text{words}}.$$

The score does not correspond to grade levels. Instead, the higher the value, the easier the text is considered to be. A text with a score of 100 should be easily understood by 11-year-old students, while a text with a score of 0 should be intended for university graduates.

– **Flesch–Kincaid grade level** (Kincaid et al. 1975) is similar to Flesch reading ease, but does correspond to grade levels. It is calculated as:

$$\mathrm{FKGL} = 0.39\frac{\text{words}}{\text{sentences}} + 11.8\frac{\text{syllables}}{\text{words}} - 15.59\,.$$

– **Dale–Chall readability formula** (Dale and Chall 1948) is calculated as:

$$\mathrm{DCRF} = 0.1579\frac{\text{difficult words}}{\text{words}} + 0.0496\frac{\text{words}}{\text{sentences}}\,.$$

The formula requires a predefined list of common (easy) words and the words which are not on the list are considered as difficult. The novelty of the Dale-Chall Formula was that it did not use word-length counts but a count of "hard" words which do not appear on a specially designed list of common words. This list was defined as the words familiar to most of the 4th-grade students: when 80 percent of the fourth--graders indicated that they knew a word, the word was added to the list.
Higher scores indicate that the text is harder, but the resulting score does not correspond to grade levels, nor is it appropriate for text aimed at children below 4th grade. In our analysis, we obtained the difficult words in two ways:

1. By constructing a list of "easy" words and considering every word not on the list as difficult. The list of easy words is described later in the paper.
2. By considering words with more than seven characters as difficult.

– **Spache readability formula** (Spache 1953) is calculated as:

$$\mathrm{SRF} = 0.141\frac{\text{words}}{\text{sentences}} + 8.6\frac{\text{unique difficult words}}{\text{unique words}} + 0.839.$$

Difficult words are defined as words that do not appear in the list of commonly used words, which is the same as the one used in the Dale–Chall readability formula. This method was specifically designed for texts targeting children up to the fourth grade and was not designed to perform well on harder text. The obtained score corresponds to grade levels.

– **Automated readability index** (Senter and Smith 1967) is calculated as:

$$\mathrm{ARI} = 4.71\frac{\text{characters}}{\text{words}} + 0.5\frac{\text{words}}{\text{sentences}} - 21.43.$$

The formula was designed so that it could be automatically captured in times when texts were written on typewriters and therefore it does not use information relating to syllables or difficult words. The obtained score corresponds to grade levels.

– **SMOG (Simple Measure of Gobbledygook)** (McLaughlin 1969) can be calculated as:

$$\mathrm{SMOG} = 1.043\sqrt{\text{difficult words}\frac{30}{\text{sentences}}} + 3.1291,$$

where difficult words are defined as words with three or more syllables. The score corresponds to grade levels.

– **LIX** (Bjornsson 1968) is calculated as:

$$\mathrm{LIX} = \frac{\text{words}}{\text{sentences}} + 100\frac{\text{long words}}{\text{words}},$$

where long words are defined as words consisting of more than six characters. LIX is the only measure we used that was not designed specifically for English but for a variety of languages. Because of this, it does not use syllables or a list of unique words. The score does not correspond to grade levels.

– **RIX** (Anderson 1983) is a simplification of LIX, and is calculated as:

$$\mathrm{RIX} = \frac{\text{long words}}{\text{sentences}}.$$

– **Coleman-Liau index** (Coleman and Liau 1975) is calculated as:

$$\mathrm{CLI} = 0.0588\mathrm{L} - 0.296\mathrm{S} - 15.8,$$

where L is the average number of letters per 100 words and S is the average number of sentences per 100 words. The obtained score corresponds to grade levels.

Language-Processing-Based Readability Criteria

The readability formulas described in the previous section use a low number of common criteria, such as the number of syllables in words or the number of words in a sentence. In our analysis, we also analyzed Slovene texts using the following additional statistics:

– percentage of long words,
– percentage of difficult words,
– percentage of verbs,
– percentage of adjectives,
– percentage of unique words,
– average sentence length.

Many of these (percentage of long words, difficult words, unique words, and average sentence length) are used as features in the readability measures described above. We evaluate them individually to determine how important each of them is for Slovene texts. The **percentage of verbs** is used because a higher number of verbs can indicate more complex sentences with multiple clauses. The **percentage of adjectives** was chosen because we assumed a higher percentage of adjectives could indicate longer, more descriptive sentences that are harder to understand.

To take into account richer morphology of Slovene and a less fixed word order compared to English, we computed two additional criteria:

– **Context of difficult words**, which is the average number of difficult words that appear in a context (i.e. the three words before or after the word) of a difficult word. Difficult words are defined as words that do not appear on the list of common words. The intuition behind this metric is that a difficult word that appears in the context of easy words is easier to understand than if it is surrounded by other difficult words since its meaning can be more easily inferred from the context.
– **Average morphological difficulty**, where we use the Slovene morphological lexicon Sloleks (Arhar Holdt 2009) to assign a "morphological difficulty" score to each word. Sloleks is a lexicon of word forms and contains frequency information for morphological variants of over 100,000 lemmas (base forms of words as defined in a dictionary). We use the relative frequency of a word variant compared to other variants of the same lemma as the morphological difficulty score.

In addition, we also calculated the number of words in each document, even if in our case, it cannot be interpreted as a criterion for determining readability since it is largely determined by the type of document. E.g., the documents belonging to the subcorpus of newspapers contain individual articles and are therefore short, while the subcorpus of computer magazines contains entire magazines which are considerably longer.

Analysis of Slovene Texts

In this section, we describe the methodology used for our analysis. In the first subsection, we describe the data sets on which we conducted our analysis. In the second subsection, we describe how we constructed the list of easy words used in some of the readability measures.

Data Sets

We created a set of subcorpora from the Gigafida reference corpus of written Slovene (Logar et al. 2012). Gigafida contains 39,427 Slovene texts released from 1990 to 2011, for a total of 1,187,002,502 words. We focused on texts published in magazines, newspapers, and books while ignoring texts collected from the internet. The texts in the Gigafida corpus are segmented into paragraphs and sentences, tokenized, and part-of-speech tagged using the Obeliks tagger (Grčar et al. 2012). We grouped the texts based on the intended audience, resulting in the following subcorpora:
– **Children's magazines** include magazines aimed at younger children (to be read independently or by their parents), namely Cicido and Ciciban.
– **Pop magazines** contain magazines aimed at the general public, namely Lisa, Gloss, and Stop.

– **Newspapers** contain general adult population newspapers, namely Delo and Dolenjski list.
– **Computer magazines** include magazines focusing on technical topics relating to computers, namely Monitor, Računalniške novice, PC & Mediji, and Moj Mikro.
– **National Assembly** includes transcriptions of sessions from the National Assembly of Slovenia.

In Table 1 we show the number of documents in each subcorpus and the average number of words per document. The subcorpus of newspapers contains the largest number of documents, while the subcorpus of text sourced from the National Assembly of Slovenia contains the fewest.

Table 1: The number of documents and the average number of words per document for each subcorpus.

Subcorpus	#docs	Avg. #words / doc
Children's magazines	125	5,488
Pop magazines	247	33,967
Newspapers	14,011	12,881
Computer magazines	163	110,875
National Assembly	35	58,841

Our hypothesis is that the readability measures will be able to distinguish texts from different subcorpora. We assume that children's magazines will be easily distinguishable from other genres that are addressing an adult population. We also suppose that general magazines are less complex than specialized magazines. The National Assembly transcripts were included as they differ from other texts in two major ways: a.) they are transcripts of spoken language and b.) they relate to a highly technical subject matter. Because of this, we were interested in how readability measures would grade them. To test our hypothesis and to determine how well each readability measure works, we analyzed texts from each subcorpus to obtain a score distribution for each measure. The scores were calculated separately for each source text (e.g., one magazine article, a newspaper, or one assembly session).

List of Common Words

For designing the list of common words, we took a corpus-based approach. Note that the methodology to create a list of common words from language corpora was already tested for other languages, (see e.g., Kilgarriff et al. 2014). We used four corpora to create a list of common words: Kres, Janes, Gos, and Šolar:

– **Šolar** (Kosem et al. 2011) contains 2,703 texts written by pupils in Slovenia from grades 6 to 13 (grade 6 to 9 in primary school, and grade 1 to 4 in secondary school).

The texts include essays, summaries, and answers to examination questions.

- **Gos** (Verdonik et al. 2011) contains around 120 hours of recorded spoken Slovene (1,035,101 words), as well as transcriptions of the recordings. The recordings are collected from a variety of sources, including conversations, television, radio, and phone calls. Around 10% of the corpus consists of recorded lessons in primary and secondary schools.
- **Janes** (Fišer et al. 2014) contains Slovene texts from various internet sources, such as tweets, forum posts, blogs, comments, and Wikipedia talk pages.
- **Kres** (Logar Berginc and Šuster 2009) is a sub-corpus of Gigafida that is balanced with respect to the source (e.g. newspapers, magazines, or internet).

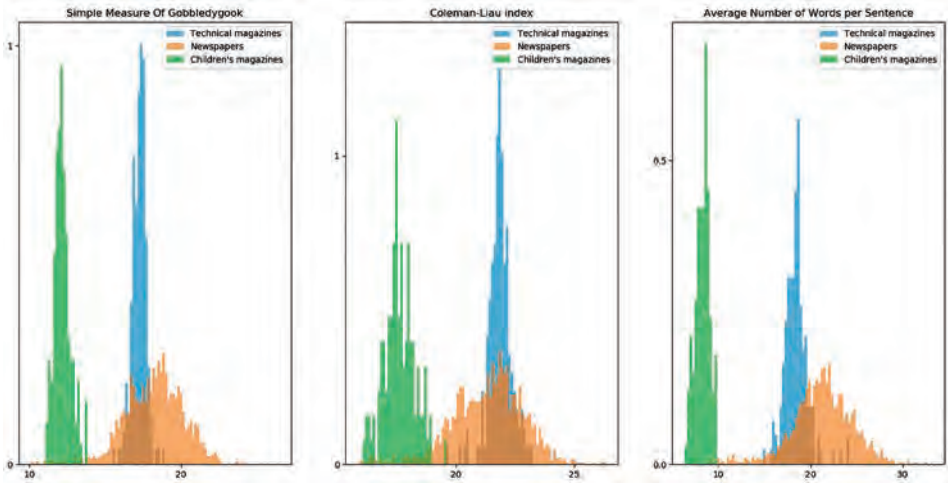
We extracted the most common words and defined the common words as the ones that appear frequently in all four corpora (and are therefore not specific to a certain text type). We use four corpora to include texts that primarily reflect language production by different language users (Gos, Janes, Šolar), as well as texts that primarily reflect standard language (Kres). We aimed at covering younger school-going population (Šolar) and adults. For some corpora, we could have assigned words to different age levels (e.g. using pupils' grade levels in Šolar or using the age groups available in Gos metadata), but these corpora are very specific and the resulting word groups would mainly reflect the genre instead of age levels. Because of this, we opted for the approach of crossing the word lists to obtain a single list. The overlap of the most common words in four corpora eliminates frequent words which are typical for only one of the corpora (e.g. administrative language in Kres, spoken language markers in Gos, Twitter-specific usage in Janes, and literary references from essays in Šolar).

From each corpus, we extracted the 10,000 most frequent word lemmas and part-of-speech tuples. In order to construct a list of common words representative of Slovene language, we selected the word lemmas that occurred in the most frequent word lists of all the four corpora. We obtained a list of 2,562 common words, which we used in readability measures.

Results

For each text in each subcorpus, we calculated readability scores using all readability measures described in the previous section. In Figure 1 we present a few examples of obtained score distributions. We show distributions for three text subcorpora (children's magazines, newspapers, and technical magazines) and three readability scores (Goobledygook, Coleman-Liau, and the average number of words in a sentence).

Figure 1: The score distributions for three text subcorpora and three readability measures. The distributions show that technical magazines readability scores are the most consistent, while newspapers' scores are more diverse. Children's magazines' scores have a strong peak on the left-hand side (easier texts) that is well separated from the other sources.



To show a compact overview of all included readability measures we calculated the median, first and third quartiles of the distribution for each score and each text subcorpus. The box-and-whiskers plots showing these results are visualized in Figure 2 which shows that most readability measures are able to distinguish between different subcorpora. Additionally, some of the readability measures confirm our original hypothesis, i.e. they are able to distinguish children's magazines from other genres that are addressing adult population, and evaluate general magazines as less complex than computer magazines.

Figure 2: The scores of each readability measure for each subcorpus of texts, represented with box plots. The subcorpora depicted from left to right are: 1.) Children's magazines, 2.) General magazines, 3.) Newspapers, 4.) Computer magazines, and 5.) National assembly transcriptions. The boxes show the first, second, and third quartile of the distributions while the whiskers extend for 1.5 IQR past the first and third quartile.

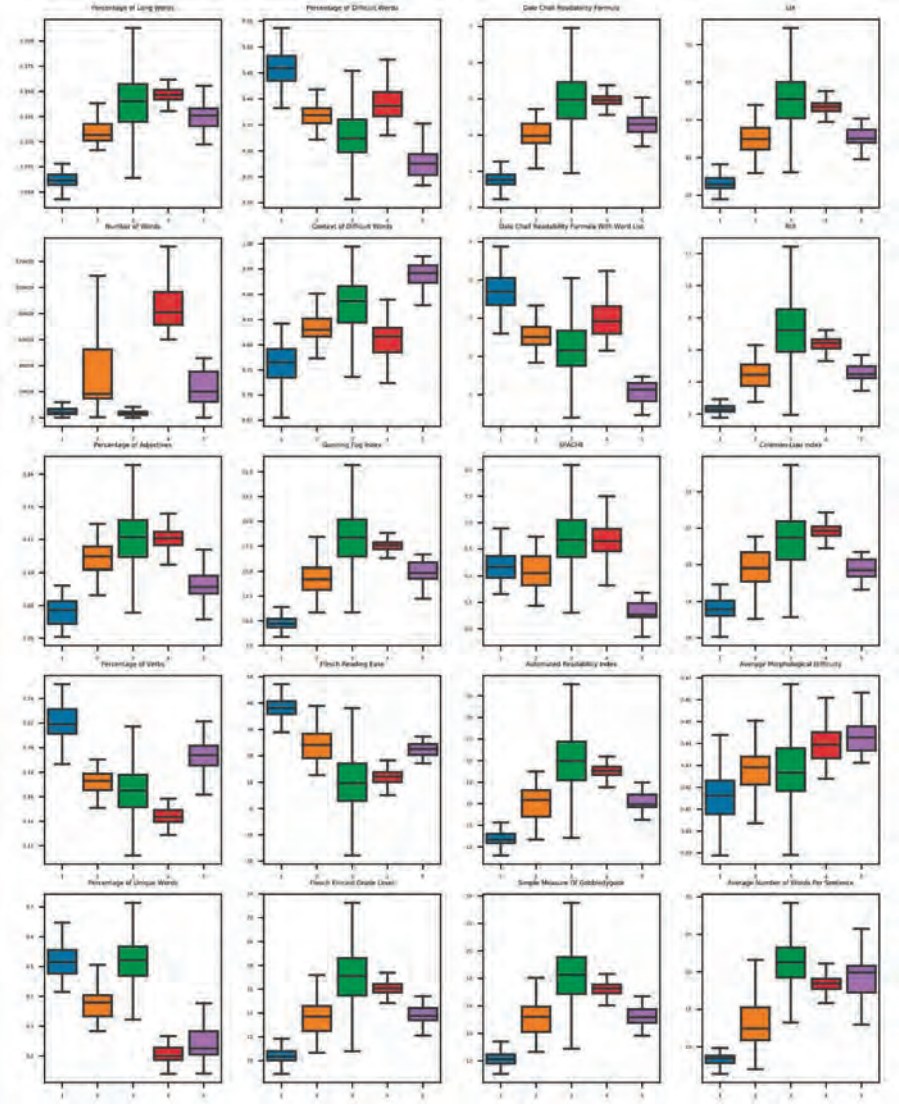


Figure 2 allows for an additional interpretation of readability measures. For example, children's magazines vs. general magazines vs. newspapers mean scores show increasing complexity in the following measures: Percentage of long words, Flesh Kincaid Grade Level, Gunning Fog Index, Dale-Chall Readability Formula (based on complexity defined by syllables), Context of Difficult Words, SMOG, LIX, RIX

and Automated Readability Index. All these measures consider the length of words and/or sentences. The percentage of adjectives also seems to correlate with the complexity of these three text types, although to a lesser extent. The same holds for Flesh Reading Ease, since higher scores indicate lower complexity. For the majority of these measures, the distinction between newspapers and specialized computer magazines is either less evident or not evident at all, but they do indicate that computer magazines are less readable than general magazines.

Scores using the list of common words do not lead to the same conclusions. Percentage of Difficult Words and Dale-Chall Readability Formula with word list do not reflect the complexity of genres, but to some extent, they do distinguish between general and specialized texts (i.e. newspapers and general magazines have lower scores than specialized computer magazines). One of the reasons for the relatively high scores for the complexity of children magazines might be in the large proportion of literary language, such as in poems for children with many words not in the list of common words. For example, "KRAH, KRAH, KRAH! MENE NIČ NI STRAH!" (Krah, krah, krah! I am not afraid!) has 7 words, out of which 4 are on the list of simple words, while the interjection KRAH is not on the simple words list. Therefore, the proportion of difficult words in this segment is 42.8% (3 occurrences of word KRAH out of 7 words in total). On the other hand, the words are short, therefore length-based measures consider them to be simple words.

The readability scores for the National Assembly subcorpus show high variability across the measures, which might be attributed to the fact that it is a different genre (spoken, but specialized). E.g., in several measures where the readability complexity rises from children's magazines to general magazines and newspapers, the National assembly scores are close to general magazines. Very long words are less likely used in spoken language, even in a political context. Average morphological difficulty and context of difficult words lead to the interpretation that this genre is more complex (less "readable"). The very high score for the context of difficult words might be attributed to enumeration of Assembly members (e.g., "Obveščen sem, da so zadržani in se današnje seje ne morejo udeležiti naslednje poslanke in poslanci: Ciril Pucko, Franc Kangler, Vincencij Demšar, Branko Kalalemina, …" (I was informed that the following deputies are occupied and cannot attend this session: …). The relatively high percentage of verbs can also be interpreted from this perspective, e.g., the National assembly text include many performatives, such as "Pričenjam nadaljevanje seje" (Starting the continuation of the session) and "Ugotavljamo prisotnost v dvorani" (Establishing the presence).

In summary, using a list of common words did not improve the partitioning of the text subcorpora perceived as easy and as difficult to read. Both measures that use it (Dale-Chall and Spache readability formulas) are poor separators. A number of simple readability measures worked well, such as the percentage of long words, the percentage of verbs/adjectives, and the average morphological difficulty.

We also calculated the sample mean and standard deviation of readability measures for each text subcorpus. The results are shown in Table 2.

Table 2: The mean and standard deviation for each subcorpus of texts and each readability score.

Measure	Children's mag.	Magazines	Newspapers	Technical mag.	National assembly
% long words	0.065 (0.015)	0.109 (0.011)	0.137 (0.029)	0.146 (0.010)	0.137 (0.046)
Number of words	5488 (6184)	33966 (34821)	12881 (84708)	110875 (151007)	58841 (106515)
% adjectives	0.078 (0.016)	0.111 (0.013)	0.120 (0.020)	0.120 (0.008)	0.096 (0.022)
% verbs	0.216 (0.026)	0.170 (0.015)	0.161 (0.034)	0.144 (0.013)	0.180 (0.044)
% unique words	0.517 (0.077)	0.375 (0.053)	0.513 (0.114)	0.244 (0.144)	0.277 (0.173)
Context of difficult words	0.756 (0.054)	0.834 (0.027)	0.849 (0.133)	0.808 (0.036)	0.929 (0.044)
% difficult words	0.464 (0.048)	0.369 (0.022)	0.356 (0.122)	0.389 (0.032)	0.280 (0.036)
Gunning Fog Index	9.950 (1.255)	14.272 (1.271)	18.662 (9.319)	17.470 (0.800)	15.901 (3.493)
Flesch reading ease	37.592 (4.989)	23.855 (5.217)	10.002 (24.128)	12.520 (4.340)	19.178 (13.098)
Flesch–Kincaid grade level	10.500 (0.894)	13.596 (1.193)	17.356 (8.959)	15.999 (0.741)	14.523 (2.761)
Dale–Chall	2.845 (0.425)	4.036 (0.306)	4.972 (1.270)	4.941 (0.258)	4.560 (0.971)
Dale–Chall with word list	7.781 (0.720)	6.534 (0.357)	6.643 (2.163)	6.955 (0.484)	5.208 (0.539)
Spache readability formula	6.217 (0.368)	6.079 (0.348)	6.977 (3.499)	6.685 (0.323)	5.482 (0.600)
Automated readability index	12.873 (1.086)	16.117 (1.428)	20.474 (11.456)	19.007 (0.885)	17.014 (3.371)
SMOG	12.206 (0.759)	15.095 (1.066)	18.200 (2.757)	17.194 (0.611)	15.849 (2.500)
LIX	33.676 (3.384)	44.999 (3.282)	56.016 (23.123)	53.260 (2.077)	47.909 (9.073)
RIX	2.381 (0.496)	4.481 (0.781)	7.370 (3.836)	6.354 (0.518)	5.250 (2.574)
Coleman-Liau index	17.785 (1.120)	19.823 (0.861)	21.220 (1.807)	21.762 (0.903)	20.318 (2.170)
Avg. morphological difficulty	0.419 (0.017)	0.428 (0.010)	0.436 (0.044)	0.441 (0.017)	0.445 (0.026)
Avg. sentence length	8.353 (0.820)	13.389 (2.843)	21.120 (4.043)	18.641 (1.960)	19.063 (3.826)

Using these results, we calculated the Bhattacharyya distance between the distributions of Children's magazines and newspapers for each score. The Bhattacharyya distance measures the similarity between two statistical distributions. We assumed the scores were distributed normally, as the results shown in Figure 1 show that the scores approximately follow a normal distribution, and calculated the distance using the following formula:

$$D_B(p,q) = \frac{1}{4}\ln\left[\frac{1}{4}\left(\frac{\sigma_p^2}{\sigma_q^2}+\frac{\sigma_q^2}{\sigma_p^2}+2\right)\right]+\frac{1}{4}\left(\frac{(\mu_p-\mu_q)^2}{\sigma_p^2+\sigma_q^2}\right)$$

We also show the Bhattacharyya coefficient, which measures the overlap between two statistical distributions and can be calculated as:

$$BC(p,q) = e^{(-D_B(p,q))}$$

The results are presented in Table 3. These results are similar to the ones shown in Figure 2, with the readability formulas using the list of difficult words showing less dichotomization power. The largest distance is obtained using average sentence lengths.

Table 3: The Bhattacharyya distances and coefficients between the distributions of scores for children's magazines and newspapers for each readability measure. The results are sorted by decreasing distance.

Measure	Distance	Coefficient
Average sentence length	2.866	0.057
SMOG	1.433	0.239
% long words	1.350	0.259
RIX	1.101	0.333
Flesch-Kincaid grade level	0.956	0.385
Automated readability index	0.945	0.389
Dale-Chall readability formula	0.885	0.413
Gunning fog index	0.880	0.415
LIX	0.853	0.426
Spache readability formula	0.797	0.451
Flesch reading ease	0.776	0.460
% adjectives	0.719	0.487
Coleman-Liau index	0.708	0.493
% verbs	0.432	0.649
% difficult words	0.365	0.694
Dale-Chall with word list	0.318	0.728
Context of difficult words	0.285	0.752
Avg. morphological difficulty	0.235	0.790
% unique words	0.039	0.961

Additional Statistical Tests

In addition to the initial analysis presented in the previous section, we performed additional, more thorough statistical tests to determine which of the evaluated measures are better at predicting the group a text belongs to. We used the following approaches:

– **Mutual information.** This measure reports the amount of information we get about a random variable *Y* by observing another random variable *X*. In our case, mutual information reports the amount of information we get about the group of texts by knowing a score of certain readability measure. Mutual information is defined as:

$$\sum_{y \in Y} \sum_{x \in X} p(x,y) log \left(\frac{p(x,y)}{p(x)p(y)} \right),$$

where p(x) and p(y) are the marginal probability distribution functions of *X* and *Y* and p(x, y) is the joint probability function of *X* and *Y*. In our case, X represents the distributions of readability measures and Y the distribution of groups. The higher the mutual information between the readability measure and the groups, the more useful the measure for determining the group membership.

– **Analysis of variance (ANOVA).** This measure first splits samples of a statistical distribution into several groups (in our case, based on the group the texts belong to) and then calculates if the groups are significantly different from one another. We use this measure to determine if the distributions obtained by calculating a single measure on each group of texts are significantly different. If they are, they can be useful for determining the group membership of a given text.

– **Feature selection using a chi-squared test.** Similarly to mutual information, we use the chi-squared test to determine whether the readability measures and the group memberships are mutually dependent. If they are, this indicates that knowing the value of the readability measure is useful when determining which group a text belongs to.

In addition to the four statistical tests used above, we also ranked each feature using a random forest classifier (Breiman 2001). The classifier is capable of automatically combining different readability measures in order to predict which subcorpus a given text belongs to and is also capable of calculating how important each readability measure was when making the prediction. The classifier is described in more detail in the next section. Using each of these tests, we obtained scores that tell us how useful each readability measure is when trying to predict the subcorpus it came from. The results are presented in Table 4, with higher scores indicating better (more informative) readability measures.

Table 4: The ranks of readability measures obtained by the statistical tests, which report the usefulness of readability measures for predicting group membership. The measures are ordered from the most useful to the least useful.

Random Forest	ANOVA	Mutual information	Chi2
Average sentence length	Average sentence length	Average sentence length	% new words
% new words	% difficult words SPG	RIX	Number of words
Number of words	% long words	SMOG	% unique words
% unique words	SMOG	Percentage of new words	Flesch reading ease
% difficult words SPG	Dale-Chall	Automated readability index	LIX
Gunning fog index	Percentage of adjectives	Gunning fog index	Average sentence length
Percentage of verbs	Coleman-Liau index	LIX	% difficult words
RIX	Percentage of unique words	Number of words	Gunning fog index
Dale-Chall (word list)	RIX	Flesch-Kincaid grade level	Automated readability index
SMOG	% verbs	Flesch reading ease	% difficult words SPG
LIX	Flesch reading ease	Dale-Chall	Flesch-Kincaid grade level
Flesch-Kincaid grade level	Context of difficult words	% unique words	SMOG
Context of difficult words	LIX	% long words	RIX
Dale-Chall	Gunning fog index	% difficult words	Coleman-Liau index
% long words	Flesch-Kincaid grade level	% difficult words SPG	Dale-Chall
% difficult words	% difficult words	Spache readability formula	Spache readability formula
Avg morphological difficulty	Automated readability index	Context of difficult words	Dale-Chall (word list)
Automated readability index	% new words	Coleman-Liau index	% long words
% adjectives	Number of words	% verbs	Context of difficult words
Flesch reading ease	Dale-Chall (word list)	% adjectives	% verbs
Spache readability formula	Spache readability formula	Dale-Chall (word list)	% adjectives
Coleman-Liau index	Avg morphological difficulty	Avg morphological difficulty	Avg morphological difficulty

The results of the statistical tests show that the features commonly used by the readability formulas (i.e. an average sentence length and number of long words) are

useful when it comes to determining group membership. In particular, the average sentence length stands out since it is ranked as the most important measure in three out of the four tests. At least one of either LIX or RIX is also highly ranked (in the top 50% of all measures) by all the tests. Those measures are the only ones from the tested measures that were not designed specifically for English, which could be one of the reasons why they perform better on Slovene texts. The results also show that a number of proposed simpler readability criteria, such as the percentage of verbs, percentage of adjectives, and the average morphological difficulty are less useful than the established statistical formulas. The results are inconclusive about the most useful readability criterion for Slovene. Several formulas and statistics are useful, but the rankings are different by different tests. When using our list of common words Dale-Chall and Spache readability formulas are again shown to perform worse than the formulas that consider long words as difficult.

Classification Results

In addition to statistical evaluation, we also performed a test with machine learning classifiers (Kononenko and Kukar 2007) to see whether we could use our readability measures to predict which subcorpus a text belongs to. With classification models, we can automatically learn how to split the texts into different subcorpora based on readability formulas and other readability criteria. We used the following classification models.

- **Decision trees** construct a binary decision tree where each node splits the training set based on one readability measure. The trained tree can predict the subcorpus of a given text.
- **Random forests (Breiman 2001)** create multiple decision trees in a random manner. This reduces the variance of a model and often gives better prediction accuracy than using a single decision tree.
- **Naive Bayes** is a probabilistic model based on the Bayes' theorem. The model assumes that the readability measures are independent.
- **Extreme gradient boosting (Chen and Carlos 2016)** constructs a large number of simple classifiers and combines them to achieve state-of-the-art results on many classification problems.

In order to use classification models, we first train them on a training subset of our data set. We used randomly selected 75% of our data set for the training. To evaluate the models, we calculated the classification accuracy (i.e. the percentage of texts each model predicted correctly) on the remaining 25% of the data set. The obtained results are presented in Table 5. The results obtained by the majority classifier (i.e. classifying everything as the most frequent group) are presented as a baseline score.

Table 5: The classification accuracies for each of the models. The numbers show the percentage of texts for which the group membership was correctly predicted.

Model	Classification Accuracy
Random Forest	0.984
Extreme Gradient Boosting	0.979
Decision Tree	0.960
Majority Classifier	0.791
Naive Bayes	0.553

Table 5 shows that we are able to predict the correct group of a text with high accuracy, over 98% with the best-performing model (Random forest). This shows that a combination of readability measures that we evaluated in this paper can be used to accurately distinguish between different groups of text.

Conclusion and Future Work

We analyzed statistical distributions of well-known readability measures on Slovene texts. We extracted five subcorpora of texts from the Gigafida corpus with commonly perceived different readability levels: children magazines, popular magazines, newspapers, technical magazines, and national assembly texts. We find that the readability formulas are able to distinguish between these subcorpora reasonably well, with the exception of national assembly texts, which are of a different, spoken, genre and the used measures were not originally designed to handle it. A number of simple readability statistics, such as the context of difficult words and average sentence length, also dichotomize the different subcorpora of text.

In this work, we only focused on simple readability formulas along with some additional readability criteria. There exist several more complex methods for evaluating the complexity of texts, such as the one presented in Lu (2009) and Wiersma et al. (2010). Such advanced methods might be more suitable for Slovene texts than the simple methods used in this paper, and we plan to test them in future work.

Most of the used English readability formulas were designed to correlate with school grades and were initially tuned on that domain. For Slovene, there currently is no publicly available data set with texts tagged according to the appropriate grade level. This disallows analysis of the readability measures from this perspective. In future work, we plan to prepare such a corpus and design several readability scores fit for different purposes. This will allow us to frame text complexity as a classification problem with the goal of predicting the grade level of a text instead of predicting its group membership. In a similar approach, experts would annotate texts with readability scores. This would allow us to fit a regression model using the readability measures analyzed in this paper.

Another area that we plan to explore is the use of coherence and cohesion measures (Barzilay and Lapata 2008; Crossley et al. 2016), which are used to determine if words, sentences, and paragraphs are logically connected. Coherence and cohesion methods usually use machine learning approaches that mostly rely on language-specific features and shall be therefore evaluated on Slovene texts. The same applies to readability measures based on machine learning (Francois and Miltsakaki 2012) which we also plan to analyze in the future.

Acknowledgments

The research was financially supported by the Slovenian Research Agency through project J6-8256 (New grammar of contemporary standard Slovene: sources and methods), project J5-7387 (Influence of formal and informal corporate communications on capital markets), a young researcher grant, research core fundings no. P6-0411 and P2-0103; Republic of Slovenia, Ministry of Education, Science and Sport/European social fund/European fund for regional development/European cohesion fund (project Quality of Slovene textbooks, KaUč). This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153 (EMBEDDIA).

Sources and Literature

Literature:

- Anderson, Jonathan. 1983. "LIX and RIX: Variations on a Little-known Readability Index." *Journal of Reading* 26, No. 6: 490–96.
- Arhar Holdt, Špela. 2009. "Učni korpus SSJ in leksikon besednih oblik za slovenščino." *Jezik in slovstvo* 54, No. 3–4: 43–56.
- Bailin, Alan, and Ann Grafstein. 2016. *Readability: Text and context*. Springer.
- Barzilay, Regina, and Mirella Lapata. 2008. "Modeling Local Coherence: An Entity-based Approach." *Computational Linguistics* 34, No. 1: 1–34.
- Björnsson, Carl Hugo. 1968. *Läsbarhet*. Liber.
- Breiman, Leo. 2001. "Random forests." *Machine learning* 45, No. 1: 5–32.
- Chen, Tianqi, and Carlos Guestrin. 2016. "Xgboost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. ACM.
- Coleman, Meri, and Ta Lin Liau. 1975. "A Computer Readability Formula Designed for Machine Scoring." *Journal of Applied Psychology* 60, No. 2: 283.
- Crossley, Scott A., Kristopher Kyle, and Danielle S. McNamara. 2016. "The tool for the automatic analysis of text cohesion (TAACO): Automatic Assessment of Local, Global, and Text Cohesion." *Behavior Research Methods* 48, No. 4: 1227–37.
- Dale, Edgar, and Jeanne S. Chall. 1948. "A Formula for Predicting Readability: Instructions." *Educational Research Bulletin*: 37–54.

- Dębowski, Łukasz, Bartosz Broda, Bartłomiej Nitoń, and Edyta Charzyńska. 2015. "Jasnopis–A Program to Compute Readability of Texts in Polish Based on Psycholinguistic Research." In *Natural Language Processing and Cognitive Science*, edited by B. Sharp, W Lubaszewski and R. Delmonte, 51–61. Liberia Editrice Cafoscarina.
- Fišer, Darja, Tomaž Erjavec, Ana Zwitter Vitez, and Nikola Ljubešić. 2014. "JANES se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino." In *Language technologies : proceedings of the 17th International Multiconference Information Society - IS 2014*, edited by Tomaž Erjavec and Jerneja Žganec Gros, 56–61. Ljubljana: Jožef Stefan Institute.
- François, Thomas, and Eleni Miltsakaki. 2012. "Do NLP and Machine Learning Improve Traditional Readability Formulas?" In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, edited by Sandra Williams, Advaith Siddharthan and Ani Nenkova, 49–57. Association for Computational Linguistics.
- Grčar, Miha, Simon Krek, and Kaja Dobrovoljc. 2012. "Obeliks: statisticni oblikoskladenjski oznacevalnik in lematizator za slovenski jezik." In *Proceedings of the Eighth Language Technologies Conference,* edited by Tomaž Erjavec and Jerneja Žganec Gros, 89–94. Ljubljana: Jožef Stefan Institute.
- Gunning, Robert. 1952. *The technique of clear writing*. McGraw-Hill.
- Justin, J. 2003. *Učbenik kot dejavnik uspešnosti kurikularne prenove: poročilo o rezultatih evalvacijske študije.*
- Kilgarriff, Adam, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. 2014. "Corpus-based Vocabulary Lists for Language Learners for Nine Languages." *Language Resources and Evaluation* 48, No. 1: 121–63.
- Kincaid, J. Peter, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease formula) for navy enlisted personnel*. Report No. 8–75.
- Kononenko, Igor, and Matjaž Kukar. 2007. *Machine Learning and Data Mining*. Chichester, Horwood Publishing.
- Kosem, Iztok, Tadeja Rozman, and Mojca Stritar. 2011. "How do Slovenian Primary and Secondary School Students Write and What Their Teachers Correct: A Corpus of Student Writing." In *Proceedings of Corpus Linguistics Conference 2011, ICC Birmingham*, 20–22.
- Logar Berginc, Nataša, and Simon Šuster. 2009. "Gradnja novega korpusa slovenščine." *Jezik in slovstvo* 54: 57–68.
- Logar Berginc, Nataša, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt, Simon Krek, and Iztok Kosem. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko and Faculty of Social Sciences.
- Lu, Xiaofei. 2009. "Automatic Measurement of Syntactic Complexity in Child Language Acquisition." *International Journal of Corpus Linguistics* 14, No. 1: 3–28.
- Mc Laughlin, G. Harry. 1969. "SMOG Grading - a New Readability Formula." *Journal of Reading* 12, No. 8: 639–46.
- Senter, R. J., and Edgar A. Smith. 1967. *Automated Readability Index*. Ohio; University of Cincinnati.
- Sherman, Lucius Adelno. 1893. *Analytics of Literature: A Manual for the Objective Study of English Prose and Poetry*. Boston: Ginn.
- Škvorc, Tadej, Simon Krek, Senja Pollak, Špela Arhar Holdt, and Marko Robnik-Šikonja. 2018. "Evaluation of Statistical Readability Measures on Slovene Texts." In *Proceedings of the conference on Language Technologies & Digital Humanities 2018,* edited by Darja Fišer and Andrej Pančur, 240–47. Ljubljana: Ljubljana University Press, Faculty of Arts.
- Spache, George. 1953. "A New Readability Formula for Primary-grade Reading Materials." *The Elementary School Journal* 53, No. 7: 410–13.

- Verdonik, Darinka, Ana Zwitter Vitez, and Hotimir Tivadar. 2011. *Slovenski govorni korpus Gos*. Trojina, zavod za uporabno slovenistiko.
- Wiersma, Wybo, John Nerbonne, and Timo Lauttamus. 2010. "Automatically Extracting Typical Syntactic Differences from Corpora." *Literary and Linguistic Computing* 26, No. 1: 107–24.
- Zwitter Vitez, Ana. 2014. "Ugotavljanje avtorstva besedil: primer »Trenirkarjev«." In *zbornik Devete konference Jezikovne Tehnologije Informacijska družba – IS*, edited by Tomaž Erjavec and Jerneja Žganec Gros, 131–34. Ljubljana: Jožef Stefan Institute.

Tadej Škvorc, Simon Krek, Senja Pollak, Špela Arhar Holdt, Marko Robnik-Šikonja

PREDICTING SLOVENE TEXT COMPLEXITY USING READABILITY MEASURES

SUMMARY

In English, the problem of determining text readability (i.e. how easy a text is to understand) has long been a topic of research, with its origins in the 19th century. Since then, many different methods and readability measures have been developed, often with the goal of determining whether a text is too difficult for its target age group. Even though the question of readability is complex from a linguistic standpoint, a large majority of existing measures are based on simple heuristics. Since most of these measures were developed for English texts, it is hard to say how well they would perform on Slovene texts. Measures designed for English are designed to correspond with the American school system, are sometimes based on pre-constructed lists of easy words which do not exist for Slovene and do not take into account morphological information when determining whether a word is difficult or not.

In our work, we analyze some common readability measures on Slovene text. We also introduce and analyze two additional readability criteria that do not appear in any of the analyzed readability measures: **morphological difficulty**, where we assume word forms that appear rarely are harder to understand than the ones that appear commonly and the **context of difficult words**, where we assume difficult words are easier to understand in a context of simple words, as their meaning can be inferred from that context. We performed the analysis on 14,581 text documents from the Gigafida corpus, which were split into five groups based on their target audience (childrens' magazines, pop magazines, newspaper articles, computer magazines, and transcriptions of sessions of the National Assembly). We assumed that the groups should have different readability scores due to their differing target audiences and writing styles.

For each analyzed readability measure we checked how well it separates texts from different groups. We did this by first obtaining the statistical distribution of readability

scores for texts in each group and checking how much the distributions differ. We show that a number of common readability measures designed for English work well on Slovene texts. To determine which of the measures perform the best we used several statistical tests.

We also show that machine-learning methods can be used to accurately (over 98% chance of a correct prediction) predict which group a text belongs to based on its readability scores. We trained four different machine-learning models (decision trees, random forests, naïve Bayes classifier, and extreme gradient boosting) and evaluated them on our dataset. We obtained the best result (98.4% classification accuracy) by using random forests.

Tadej Škvorc, Simon Krek, Senja Pollak, Špela Arhar Holdt, Marko Robnik-Šikonja

NAPOVEDOVANJE KOMPLEKSNOSTI SLOVENSKIH BESEDIL Z UPORABO MER BERLJIVOSTI

POVZETEK

Problem berljivosti (t.j. kako enostavno je besedilo za branje) je v angleščini dobro raziskan. Obstaja veliko različnih metod in formul, s katerimi lahko analiziramo angleška besedila z vidika berljivosti. Kljub temu, da je vprašanje berljivosti z lingvističnega vidika zapleteno večina metod za ugotavljanje berljivosti temelji na preprostih značilnostih besedil. Ker je bila večina mer berljivosti zasnovanih za angleška besedila, ne moremo biti prepričani da bodo enako dobro delovala na slovenskih besedilih. Angleške mere berljivosti so namreč usklajene z ameriškim šolskim sistemom, včasih temeljijo na vnaprej sestavljenih seznamih lahkih besed in ne upoštevajo težavnosti besed z morfološkega vidika.

V našem delu analiziramo pogoste mere berljivosti na slovenskih besedilih. Poleg tega uvedemo in analiziramo dva dodatna kazalnika berljivosti ki ne nastopata v pogostih merah berljivosti: **morfološka zahtevnost besed**, s katero želimo zajeti predpostavko da so redkejše morfološke oblike besed težko berljive, in **kontekst težkih besed**, s katero želimo zajeti predpostavko, da so neznane besede, ki se pojavijo v kontekstu znanih besed lažje berljive, saj lahko njihov pomen razberemo iz konteksta. Analizo smo izvedli na 14,581 besedilih iz korpusa Gigafida, ki smo jih razdelili v pet skupin glede na njihovo ciljno publiko (Otroške revije, splošne revije, časopisni članki, računalniške revije in transkripcije sej Državnega zbora). Predpostavili smo, da imajo revije zaradi različnih ciljnih publik in tematik različne sloge pisanja in posledično različne stopnje berljivosti.

Za vsako izmed mer berljivosti smo preverili, kako dobro med seboj loči besedila iz različnih skupin. Za vsako izmed njih smo pridobili statistično distribucijo vrednosti berljivosti vsake skupine in preverili, ali so distribucije ustrezno ločene. V analizi pokažemo, da se številne uveljavljene mere, ki so bile zasnovane za angleščino, dobro obnesejo tudi na slovenskih besedilih. Da bi ugotovili, katere mere najbolje razlikujejo med skupinami smo uporabili statistične teste.

Poleg tega pokažemo, da lahko z modeli strojnega učenja in kombinacijo analiziranih metod berljivosti z visoko točnostjo (nad 98 %) napovemo, v katero skupino spada določeno besedilo. Za to analizo smo uporabili štiri različne metode strojnega učenja (odločitvena drevesa, naključne gozdove, naivni Bayesov klasifikator, in extreme gradient boosting). Najboljši rezultat (98,4 %) smo dobili z metodo naključnih gozdov.

Reviews and Reports

Language Technologies and Digital Humanities 2018, 20–21 September 2018, Faculty of Electrical Engineering, Ljubljana

The conference *Language Technologies and Digital Humanities 2018* took place at the Faculty of Electrical Engineering at the University of Ljubljana on 20 and 21 September 2018. It was organised by the *Slovenian Language Technologies Society,*[1] the *Centre for Language Resources,*[2] the *Faculty of Electrical Engineering,*[3] and the research infrastructures *CLARIN.SI*[4] and *DARIAH.SI.*[5] The conference was the eleventh iteration – as well as the 20th anniversary – of the *Language Technologies* conference series,[6] which was started by the *Slovenian Language Technologies Society* and has been taking place biennually since 1998. In 2016 it successfully expanded its scope to include Digital Humanities as well. The 2018 edition of the conference was very international, with authors from 17 European countries[7] as well as two participants from Brazil and Japan. This is why the conference programme was organized in such a way that talks on Day 1 were in English and on Day 2 in Slovene.

The conference was opened by the first keynote speaker Malvina Nissim, who is Associate Professor of Computational Linguistics and Natural Language Processing at the University of Groningen. In her talk, titled "Too good to be true: Current Approaches to author profiling", she discussed novel approaches to the automatic identification of the gender and age of social media users. In particular, she showed that models which abstract away from the lexical content of social media posts and instead focus on extra-linguistic information such as punctuation and emoticons, whose use is shared across languages to a great extent, offer a robust and reliable way to identify such personal information.

1 *SDJT – Slovensko društvo za jezikovne tehnologije*, http://www.sdjt.si/wp/english/.
2 *CJVT – Centre for language resources and technologies*, https://www.cjvt.si/en/.
3 *Univerza v Ljubljani, Fakulteta za elektrotehniko*, http://www.fe.uni-lj.si/en/.
4 *CLARIN Slovenia*, http://www.clarin.si/info/about/.
5 *Dariah-SI | Digitalna humanistika*, http://www.dariah.si/en/.
6 *SDJT – Slovensko društvo za jezikovne tehnologije*, http://www.sdjt.si/wp/dogodki/konference/strani/.
7 In addition to Slovenia, the following European countries were represented: Austria, Belgium, Bulgaria, Denmark, Finland, Germany, Greece, Ireland, the Netherlands, Norway, Poland, Portugal, Serbia, Spain, Sweden, and Switzerland.

The keynote talk was followed by two morning sessions devoted to topics in machine translation and language resources. The machine translation session was chaired by Tomaž Erjavec and comprised two talks. Gregor Donaj and Mirjam S. Maučec compared traditional statistical machine translation with the use of neural networks for translating between Slovenian and English, while Mihael Arčan compared the two approaches by using translations between three Slavic languages – Slovenian, Croatian and Serbian.

In the subsequent session devoted to language resources, which was chaired by Simon Krek, six papers were presented, introducing on-going work on language corpora and lexical resources in Slovenian, Croatian and Portuguese. For example, Filip Dobranić presented joint work with Nikola Ljubešić, Darja Fišer and Tomaž Erjavec on the creation of the *Parlameter* corpus, which contains contemporary Slovenian parliamentary proceedings from 2014 to 2018 with rich speaker metadata on the gender, age, education and party affiliation of the members of the Slovenian parliament. Filip also showcased how the resource facilitates in-depth exploration of institutionalised language use and interpersonal behaviour patterns, which is important for an interdisciplinary approach to the analysis of parliamentary discourse that involves collaboration between researchers working in disciplines like sociology, discourse analysis, history, sociolinguistics, and political science.

The poster session presented nine posters on various applications of quantitative approaches to data analysis within digital humanities and social sciences. For instance, Katja Mihurko Poniž and colleagues introduced a tool that aids in the research of the historical representation of women's authorship, which is an important topic in sociohistoric approaches to literary theory, while Damjan Popič and Darja Fišer presented a corpus-driven analysis of the attitudes toward language in Slovenian, Croatian, and Serbian computer-mediated communication.

The first afternoon session, which was chaired by Jurij Hadalin, was devoted to Digital Humanities. Dan Podjed and Ajda Pretnar analysed the use of social media by the Slovenian President Borut Pahor for self-promotion. On the basis of qualitative and quantitative approaches to data analysis, they identified three distinct categories of the President's Instagram posts that prove to be the most popular among his followers; namely, (i) photographs in which he is seen together with celebrities and his family, (ii) posts in which he gives the impression of being approachable, and (iii) photographs in which he is depicted in an unusual situation. Tobias Weber and Jeremy Bradler discussed a novel approach of integrating computational methods, digital resources and computer literacy skills into the curriculum of Finno-Ugric linguistics, stressing the importance of tailoring the materials to the students' non-computational backgrounds in humanities and social sciences.

The subsequent session, which was chaired by Simon Dobrišek, concluded the first day of the conference. Nine papers were presented on topics related to language technologies and their application. For instance, in a cross-disciplinary approach to phonetics and medicine, Tatjana Marvin introduced joint work with Jure Derganc,

Samo Beguš and Saba Battelino on a novel Slovenian Sentence Matrix Text for measuring speech intelligibility in patients suffering from hearing loss. To give another example in a different field of application, Milan van Lange and Ralf Futselaar presented their use of word embeddings in the analysis of parliamentary debates on war criminals in The Netherlands.

The second day of the conference began with a keynote talk delivered by Martijn Kleppe, who is Head of the Research Department at the National Library of the Netherlands. In his talk, titled "Bringing Digital Humanities to the wider public: libraries as incubator for DH research results", Martijn presented one of the main aims of the National Library of the Netherlands, which is to support researchers in the Digital Humanities and social sciences and incorporate their research results in its services and products. To this end, Martijn showcased *LAB,* and online toolchain of the Library which offers researchers an interoperative environment for working with richly annotated texts and state-of-the-art tools for processing handwritten documents. He also discussed the Institute's collaborations with other national and international research infrastructures, such as CLARIN ERIC.

The next session, chaired by Andrej Pančur, brought two talks on issues related to Slovenian research infrastructures. Maja Dolinar, Janez Štebe and Sonja Bezjak, presented a new set of guidelines for the acquisition and archiving of qualitative research in the Slovenian Social Science Data Archives. Tomaž Erjavec presented joint work with Darja Fišer and Jakob Lenardič on how linguistic data, such as those found in language corpora, are cited in Slovenian research publications, and proposed recommendations and solutions for more consistent and rigorous citation practices in line with the Austin Principles of Data Citation in Linguistics.

In the next session, chaired by Darja Fišer, six talks were given on topics related to corpus linguistics. For instance, Nataša Logar presented the main morphosyntactic characteristics of academic Slovenian, which she analysed together with Tomaž Erjavec on the basis of the Slovenian balanced corpus Kres and the corpus KAS, which consists of Slovenian BA, BSc and PhD theses. Iztok Kosem presented joint work with Simon Krek, Polona Gantar, Špela Arhar Holdt, Jaka Čibej, and Cyprian Laskowski on the user interface of the first Collocations Dictionary for Modern Slovenian, which was compiled on the basis of state-of-the-art lexicographic methods.

The student session, which was chaired by Iza Škrjanec, included four talks. Urška Bratoš discussed the compilation and analysis of a corpus of tweets written by Slovenian politicians. Isolde van Dorst then presented a statistical analysis of Shakespeare's use of pronominal expressions, specifically his usage of the second-person pronoun *you* and its two now-obsolete informal variants – nominative *thou* and accusative *thee*. Gabi Rolih presented an implementation of a K-means clustering method applied to computer-mediated communication and discussed how it can be used to further improve a state-of-the-art part-of-speech tagger for Slovenian. Finally, Klara Eva Kukovičič compared the concordancer Sketch Engine with the tool CollTerm from the point of view of terminology extraction. The Best Student Paper Award was awarded to Isolde van

Dorst by the selection committee Iza Škrjanec (chair of the Student Session), Tomaž Erjavec (on behalf of the Programme Committee) and Kaja Dobrovoljc (on behalf of the *Slovenian Language Technologies Society*).

The final session, chaired by Matija Ogrin, focused again on Digital Humanities and concluded the conference. Five papers were presented on topics related to cultural heritage, historical studies and geography. For instance, Andrej Pančur presented the SIstory web portal, which offers a sustainable repository for digital editions of historical texts, while Alenka Kavčič presented joint work with Ivan Lovrić and Vera Smole on the development of an interactive online map of the seven major Slovenian dialect groups, which includes geocoded text examples enriched with audio materials that exemplify the salient phonological features of the dialects.

The *Language Technologies and Digital Humanities 2018* conference successfully presented on-going and completed work on state-of-the-art language tools and resources, as well as their application. The presentations that used computational tools and methodologies to answer qualitative research questions were especially illustrative in showing how language technologies facilitate and open new grounds for research in fields like translation studies, political science, historical studies, phonetics and phonology, and literary theory. Perhaps crucially, the work presented by master's and doctoral students was an inspiring showcase of how young researchers use innovative computational approaches to tackle complex research problems in such interdisciplinary fields. The conference thus gave both novice and experienced researchers from Slovenia and abroad a chance to strike up collaborations and get involved in research projects that bridge the gap between language technologies on the one hand and humanities and social sciences on the other.

Jakob Lenardič*

* **Department for Translation, Faculty of Arts, University of Ljubljana, Aškerčeva 2, SI-1000 Ljubljana, jakob.lenardic@ff.uni-lj.si**

Sources and Literature

- Arčan, Mihael. 2018. "A comparison of Statistical and Neural Machine Translation for Slovene, Serbian and Croatian." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 3–10. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Bratoš, Urška. 2018. "Gradnja korpusa tvitov slovenskih politikov Janes-TwePo." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 269–73. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Dolinar, Maja, Janez Štebe, and Sonja Bezjak. 2018. "Razvoj smernic za predajo in arhiviranje kvalitativnih podatkov v Arhivu družboslovnih podatkov." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 55–61. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Donaj, Gregor, and Mirjam S. Maučec. 2018. "Prehod iz statističnega strojnega prevajanja na prevajanje z nevronskimi omrežji za jezikovni par slovenščina-angleščina." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 62–68. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- van Dorst, Isolde. 2018. "You, Thou and Thee: A Statistical Analysis of Shakespeare's Use of Pronominal Address Terms," In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 274–80. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Fišer, Darja, Jakob Lenardič, and Tomaž Erjavec. 2018. "Citiranje jezikoslovnih podatkov v slovenskih znanstvenih objavah: stanje in priporočila." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 77–84. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Kavčič, Alenka, Ivan Lovrić, and Vera Smole. 2018. "Karta slovenskih narečnih besedil." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 121–25. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Kleppe, Martijn. 2018. "Bringing Digital Humanities to the Wider Public: Libraries as Incubator for DH Research Results." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 2. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Kosem, Iztok, Simon Krek, Polona Gantar, Špela Arhar Holdt, Jaka Čibej, and Cyprian Laskowski. 2018. "Kolokacijski slovar sodobne slovenščine," In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 133–39. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Kukovičič, Klara Eva. 2018. "Uporabnost luščilnikov terminologije Sketch Engine in CollTerm z vidika (študenta) prevajalca." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 281–87. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- van Lange, Milan, and Ralf Futselaar. 2018. "Debating Evil: Using Word Embeddings to Analyze Parliamentary Debates on War Criminals in The Netherlands." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 147–53. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Ljubešić, Nikola, Darja Fišer, Tomaž Erjavec, and Filip Dobranić. 2018. "The Parlameter corpus of contemporary Slovene parliamentary proceedings." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 162–67. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Logar, Nataša, and Tomaž Erjavec. 2018. "Strokovnoznanstvena slovenščina: besednovrstne in oblikoskladenjske značilnosti." In *Proceedings of the Conference on Language Technologies & Digital*

Humanities 2018, edited by Darja Fišer and Andrej Pančur, 175–80. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.

- Marvin, Tatjana, Jure Derganc, Samo Beguš, and Saba Battelino. 2018. "Word Selection in the Slovenian Sentence Matrix Test for Speech Audiometry." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 181–87. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Mihurko Poniž, Katja, Amelia Sanz, Marie Nedregotten Sørbø, Suzan van Dijk, Viola Parente-Čapková, Narvika Bovcon, and Aleš Vaupotič. 2018. "Teaching Women Writers with NEWW Virtual Research Environment." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 254–55. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Nissim, Malvina. 2018. "Too Good to Be True: Current Approaches to Author profiling." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 1. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Pančur, Andrej. 2018. "Trajnost digitalnih izdaj: Uporaba statističnih spletnih strani na portal Zgodovina Slovenije – SIstory." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 203–10. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Podjed, Dan, and Ajda Pretnar. 2018. "Samopromocija na Instagramu: Primer predsednikovega profila." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 221–26. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Popič, Damjan, and Darja Fišer. 2018. "Odnosi do jezika v slovenski, hrvaški in srbski računalniško posredovani komunikaciji." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 256–59. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Rolih, Gabi. 2018. "K-means Clustering of CMC Data for Tagger Improvement." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 288–91. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Weber, Tobias, and Jeremy Bradley. 2018. "Exploring Finno-Ugric Linguistics Through Solving IT Problems." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 248–53. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.

THE INSTITUTE OF CONTEMPORARY HISTORY LIBRARY

The Institute of Contemporary History Library is a specialised library, collecting and storing the resources for scientific researchers and fans of contemporary history. Initially its materials mostly encompassed books and magazines on the history of World War II and history of the workers' movement. However, as the Institute's areas of interest expanded, its library has also procured materials about the political, economic, social and cultural history of Slovenians.

The library's basic collection consists of around 40.000 books about the contemporary history of Slovenia and the world. Initially the majority of books focused on the history of World War II and the workers' movement, while later the library started procuring literature about social and cultural history. We can state that with its collection of materials our library represents the most important historiographic collection about the history of the 20th century in Slovenia.

The library keeps around 200 titles of magazines, including all of the most important newspapers since Bleiweis's Kmetijske and rokodelske novice newspaper to cultural and professional magazines and all kinds of bound daily newspapers.

Opening hours: Monday – Friday: 8 a.m. to 1 p.m., Wednesday: 8 a.m. to 3 p.m.

Contact: + 386 1 200 31 28 or +386 1 200 31 32

Web page: http://www.inz.si/knjiznica.php

UDC
94(497.4)"18/19"
UDK

ISSN 0353-0329

Inštitut za novejšo zgodovino