From mouth to keyboard: the place of non-canonical written and spoken structures in lexicography

Ana Zwitter Vitez[1,2], Darja Fišer[2]

[1] Department of Applied linguistics, Faculty of Humanities,
University of Primorska, Titov trg 5, 6000 Koper
2 Department of Translation. Faculty of Arts, University of Ljubljana,
Aškerčeva 2, 1000 Ljubljana
E-mail: ana.zwitter@guest.arnes.si, darja.fiser@ff.uni-lj.si

Abstract

As user-generated content is on the rise both in terms of volume and importance, the long established relation between spoken and written communication needs to be re-examined in lexicography. This is the aim of this paper, in which we perform a corpus-based analysis of typical non-canonical words in spoken and computer-mediated communication in Slovene. The results show that the spoken and the Twitter corpus contain a similar proportion of non-standard pronunciation/spelling variants, interaction words and informal lexemes. On the opposite end of the spectrum are news comments which contain a higher proportion of nouns and a smaller proportion of non-canonical words. The presented study brings a language-independent methodology of identifying typical elements of spoken and written informal texts.

Keywords: lexicography; non-canonical language; computer-mediated communication; spoken language

1. Introduction

Contemporary corpora-based dictionaries are increasingly tackling language material from informal genres, such as tweets, forums, blogs, and comments on news portals. The stereotype of user-generated communication is that it is a hybrid between spoken and written language. Nevertheless, research shows that "netspeak is better seen as a written language which has been pulled some way in the direction of speech rather than as spoken language which has been written down" (Crystal, 2007: 47). To what extent is this true? What are the main similarities and differences between typical spoken and user-generated structures? And how should these typical structures of informal spoken and written genres be included in dictionaries? In order to attempt to answer these questions it seems reasonable to establish a methodology which enables a systematic comparison of spoken and user-generated informal communication.

This paper presents the results of a corpus-based analysis of non-canonical words in user-generated and spoken communication in Slovene. The rest of the paper is structured as follows: in Section 2 we introduce related work analysing spoken and user-generated structures in lexicography; in Section 3, we bring out the analysed

datasets; the methodological Section 4 focuses on the procedure and the main levels of analysis (part of speech, standardization, categorization, linguistic phenomena). In Section 5, we examine the results showing on which levels the analyzed subcorpora of user-generated content display the most spoken language characteristics and in the concluding section, we discuss the value of the results for Slovene and international lexicographic practices.

2. Spoken and user-generated structures in lexicography

Numerous previous studies have confirmed that "there is a whole world" (Morel, Danon Boileau, 1998) between spoken and written texts. These differences have led to the fact that spoken discourse was included in lexicography as soon as technical constraints permitted it. The first Cobuild dictionary (Sinclair et al., 1987), based on the Collins corpus, included examples of English "that people speak and write every day", including material from radio, TV and everyday conversations. Nevertheless, Moon (1998) argues that the extensive differences between written and spoken language should launch reconsideration in dictionary-making on the levels of phonology, phraseology, collocations, colligations, parts of speech and syntactic structure.

With an increasing quantity of user-generated content on the internet, the relation between spoken and written communication presents a new research challenge. Different disciplines have acknowledged the role of linguistics in the analysis of "netspeak": D. Crystal (2007) exposes sociolinguistics, stylistics, teaching, and applied linguistics. M. Beißwenger (2012) adds the importance of analysing user-generated contents for lexicography, while exposing genre-specific discourse markers and 'netspeak' jargon (like 'imho' for 'in my humble opinion'), and new vocabulary, e.g. 'funzen' (an abbreviated variant of the German verb 'funktionieren', en.: 'to function'). Due to the accessibility of user-generated texts, updating vocabulary has become a regular practice: M. Rundell (2014) reports about four updates per year in Macmillan where new words, meanings, and phrases are added (typically at a rate of around 120 to 150 per update).

In Slovene linguistics, historical, political and discipline-specific factors have promoted a protective view of the language, keeping the process of language standardisation separated from the data on actual language use (Verovnik, 2004). Monolingual lexicography is still finding its digital form (Kosem, 2015), but the prevalent doctrine of contemporary lexicography is becoming descriptive, turning away from the position of "how people 'ought to' use language" (Atkins & Rundell, 2008: 2). It therefore seems to be the right time to examine the relation between the written user-generated contents and the spoken discourse and start including user-generated contents into dictionaries.

In principle, we know what to do, but in practice, different approaches reveal

potentials and traps when trying to systematically compare spoken and user-generated communication. Linguistic studies (Akinnaso, 1982; Chovanec, 2009; Sindoni, 2013) seem to be comprehensive but are usually not based on quantitative research. On the other hand, different computational approaches give very detailed results on certain linguistic phenomena (Leech et al., 2001; Baron, 2010; Bamman et al., 2014), but only offer results on specific structures. It seems that a systematic corpus study of spoken elements in user-generated discourse could provide valuable insights and could help to resolve the dilemma of including these elements into lexicographic practice.

3. Analysed datasets

For the study presented in this paper we used three corpora:

1) a corpus of Slovene called Kres (Logar Berginc et al., 2012) which contains 100 million tokens, sampled from the reference corpus Gigafida. It contains equal proportions of literary, non-fiction, newspaper and internet texts. The corpus has been PoS-tagged and lemmatized. In our study we used it as a baseline corpus displaying canonical, standard written language use.

Example 1)

Example	**Kljub obilju, v katerem živimo, pa danes mineralov marsikomu primanjkuje, za kar je kriva nepravilna prehrana.**
Translation	*Despite the abundance in which we live nowadays, many people lack minerals, which is consequence of poor nutrition.*

2) the corpus of spoken Slovene called Gos (Verdonik & Zwitter Vitez, 2011) which contains 1 million tokens, transcribed from 120 hours of recorded spontaneous private and public speech on TV, radio, in schools, meetings, bars and at home, sampled for sex, age, region and education level of the speakers. The transcriptions were performed in two ways: one resembles speech as closely as possible while the other one is normalized in accordance with standard spelling conventions, which simplifies corpus querying but also enables the analysis of lexical variants. The transcriptions were also PoS-tagged and lemmatized. In our study we used it to identify the phenomena that are characteristic of spoken discourse.

Example 2)

Example	**pa sej itak ni nč februarja itak je eee dons je bla angleščina jutr je pa nemščina to je pa to**
Translation	*well in any case there's nothing in Febuary today we had English tomorrow we have German and that's it*

3) the corpus of Slovene user-generated content called Janes (Fišer et al., 2014) which contains 160 million tokens, collected from Twitter, forums, comments on news portals and blogs. As the corpus is rich in non-canonical lexical variants, they were standardized (Ljubešić et al., 2014) before they were PoS-tagged and lemmatized. Social media are used in two very distinct ways: as one of the official news channels by news media, government institutions, private companies and organizations who use the traditional communication conventions, and proper user-generated content in which non-professional users share their personal opinions and experience with their social network in more relaxed settings, often resorting to non-canonical communication conventions. Each text in the corpus was automatically annotated with a standardness measure at the technical and linguistic levels (Ljubešić et al., in press), making it possible to analyse only those parts of the corpus that contain non-standard language, for example.

Example 3)

Example	**a se men sam zdi al si neki našpičena dons ? : -(**
Translation	*is it just me or you really are a bit pissed off today ? : -(*

4. Methodology

The goal of the study presented in this paper was to analyse the spoken language elements in computer-mediated communication. We performed this analysis by first identifying the lexical spoken-language features with respect to standard written communication. We then compared lexical features of computer-mediated communication with traditional written communication and checked to what extent the characteristics of the user-generated contents resemble spoken language. As this is the first systematic comparison of Slovene spoken, user-generated and standard corpora, we wanted to analyse single-word units that are typical of each of the corpora. This was achieved by a three-way comparison of keyword lists (Kilgarriff et al., 2004) which were generated in the SketchEngine by comparing both the spoken-language Gos corpus and the Janes corpus of user-generated content against the Kres corpus of written Slovene. While a single keyword analysis was performed on the entire Gos corpus, three Janes subcorpora were examined separately; tweets, forum messages and news comments. We opted for an independent analysis of the three genres because we believe they display important distinctive characteristics and do not resemble spoken language in the same way and to the same degree. Since we were interested in non-canonical language phenomena, only non-standard texts (i.e. those from bands 2 and 3 of the linguistic standardness measure) were included in the analysis.

GOS		Forums		Twitter		Comments	
eee	*eee*	avto	*car*	btw	*btw*	ane	*isn't it*
mhm	*mhm*	tud	*also*	oz.	*or*	nebi	*wouldn't*
eem	*eem*	mal	*a little*	cca	*around*	nevem	*don't know*
sej	*any case*	tko	*like this*	slo	*Slovene*	ala	*like*
tud	*also*	blo	*was*	lol	*lol*	kriv	*guilty*
zdej	*now*	tut	*also*	cez	*in*	krivi	*guilty (pl.)*
tko	*like this*	gor	*up*	bos	*you will*	obsojen	*prosecuted*
aha	*oh*	jst	*I*	nic	*nothing*	fajn	*nice*
blo	*was*	mam	*have*	prevec	*too much*	cel	*whole*
tak	*like this*	gume	*tires*	mogoce	*maybe*	neprimerno	*inappropriate*

Table 1: Top 10 words from the analysed corpora[1]

The top 200 word forms were manually analysed on each of the four generated keyword lists. Each analysis consisted of four steps:

(1) Part of speech: we annotated each keyword with part-of-speech information. Since many word forms are ambiguous, we used the most frequent part of speech annotation only.

word		PoS
tko	*like this*	adverb
aha	*oh*	interjection
blo	*was*	verb

Table 2: Example of PoS annotation

(2) Standardization: First, we checked whether the keyword was canonical. If it was not, we normalized it with its standard variant. If the word form was ambiguous and could be standardized in several ways, we used the most frequent option and annotated it with a special "VARIANT" flag.

word	normalization	Translation 1	Translation 2
pol	potem_VAR	*then*	*half*

Table 3: Example of ambiguous normalization

(3) Categorization: We checked whether the keyword form was part of the standard vocabulary. If it was not, we attempted to assign them to different categories, which led us to the next 10 categories, displaying either lexical or orthographic deviations from the norm: abbreviation, omitted diacritics, discourse marker, foreign expression, informal expression, expression signalling interaction in communication, medium-specific expression, spelling resembling pronunciation, non-standard

[1] The translations into English are in italics.

tokenization and topic-specific expression. If the keyword displayed characteristics of several categories, we assigned it the most salient one.

Category	Example	Translation
pronunciation	reku	*said*
interaction	hvala	*thank you*
standard	vedno	*allways*
topic	servis	*service*
informal	folk	*people*
diacritics	cist	*totally*
medium	prijavi	*report*
tokenization	nebi	*would not*
discourse	hm	*hm*
abbreviation	cca.	*about*
foreign	good	*good*

Table 4: Categorization of the analysed keywords

(4) Linguistic phenomenon: We examined the non-canonical word forms in all 10 categories and identified the linguistic phenomenon at play in each case.

Linguistic phenomenon	Example	Translation
reduction	boljš	*better*
neutralization	dej	*come on*
from English	ful	*totally*
deixis	tale	*this*
article	ta	*the*

Table 5: Linguistic phenomenon of deviation

The results of the analysis of the spoken-language corpus and the user-generated subcorpora were compared in order to determine the degree and distribution of interference of speech/written discourse in computer-mediated communication. In the end, an analysis of the extent and distribution of orthographic variation of the non-canonical keywords found in all four analysed samples was performed.

5. Analysis and results

5.1 PoS categorization

In order to get a general picture regarding the material we are dealing with, the keywords in Gos and in user-generated corpora were annotated with part-of-speech information (Figure 1).

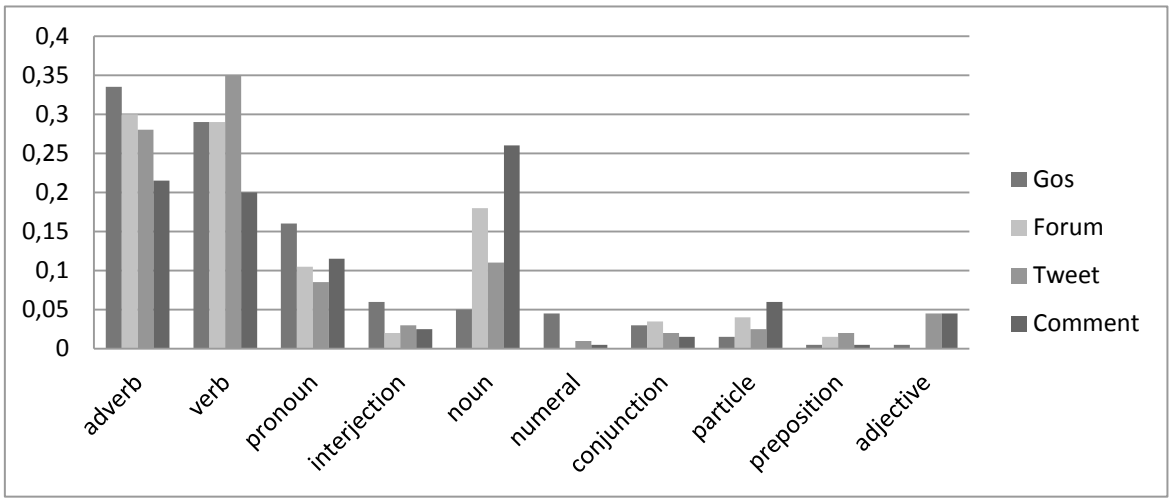

Figure 1: PoS distribution in spoken and user-generated corpora.

The results show that the most frequent PoS categories in the Gos corpus are adverb (33%), verb (29%), pronoun (16%) and interjection (6%). Within the top three typically spoken keywords we find hesitation marks *eee, mhm* and *eem* which are the consequence of simultaneous planning and uttering spoken discourse and are thus not present in the user-generated corpora. The high frequency of adverbs (e.g. *čist - totally*) is probably related to their original function of modifying other words, which helps to express the author's opinion. Numerous frequent verbs in the Gos corpus have a different pragmatic function from that assigned in the PoS process (Example 4):

Example 4)

Example	**// zakaj kako a veš mislim eee poznaš eee [ime] od prej?/**
Translation	*// why how you know I mean eee do you know [name] from before?//*

Example 4 shows that the verb *mislim* (e.g. *I think)* plays an important role in keeping attention of the addressee while formulating the rest of the utterance, so it does not function within its traditional syntactic structure (e.g. *I think that…*) but rather as a discourse marker (e.g. *I mean*).

The Forum subcorpus has a similar proportion of adverbs (30%) and verbs (29%). Many verbs relate to the expression of personal opinions or evaluations (e.g. *me zanima - I am interested, zgleda - it seems, vidim - I see*). Contrary to spoken discourse, the non-standard forum discourse is marked by frequent nouns related to the topic of conversation (e.g. *gume - tires, cena - price, poraba - consumption*) and the nature of the conversation (*problem, odgovor - answer*) where a predictable set of formulations is used, as shown in Example 5.

Example 5)

Example	**Hvala za odgovore in lep dan.**
Translation	*Thank you for you answers and have a nice day.*

The Twitter subcorpus consists of a slightly lower proportion of typical adverbs (28%) and a significantly higher proportion of verbs (35%) expressing the author's point of view (e.g. *zgleda - it seems*) or illocutionary verbs expressing promise, inquiry or request of interaction with other authors (e.g. *rabim - I need, poznam - I know, dobiš - you get*):

Example 6)

Example	**Rabim prostovoljca ki bi mi prišel skuhat mlečni riž.**
Translation	*I need a volunteer who would cook a rice pudding for me.*

The Comments corpus contains fewer verbs and adverbs but a significantly higher proportion of nouns (26%) among the top 200 analysed keywords, than the Gos corpus (only 5%). Nouns in the Comments corpus range from the emotionally marked (e.g. *sramota - shame*) to the topic-oriented (e.g. *denar - money, volitve - elections, gol - goal*):

Example 7)

Example	**Sramota. Samo to bom reku.**
Translation	*Shame. That's all I'll say.*

It is interesting to note that the process of manually annotating word class for 800 words without seeing their context is less than trivial because very often, a certain word has a traditional PoS identity but operates in a different way in the analysed corpus (this is why it would be interesting to see the score for inter-annotator agreement if many annotators were involved). This phenomenon can be shown by the example of the verb *recimo (say)* which mostly operates in the pragmatic function of a discourse connector in the Janes corpus.

5.2 Standardization

With the next level of analysis, we wanted to examine the proportion of non-canonical words among the analysed sample of 200 keywords per corpus. Within the Gos project, standardization was carried out manually (1 million words). For the Janes corpus, an automatic rudimentary standardization has been performed and added as an attribute, but it is currently too imprecise for detailed analysis. This is why we have performed the process of standardization manually for the purpose of this research following the guidelines of the Gos project.

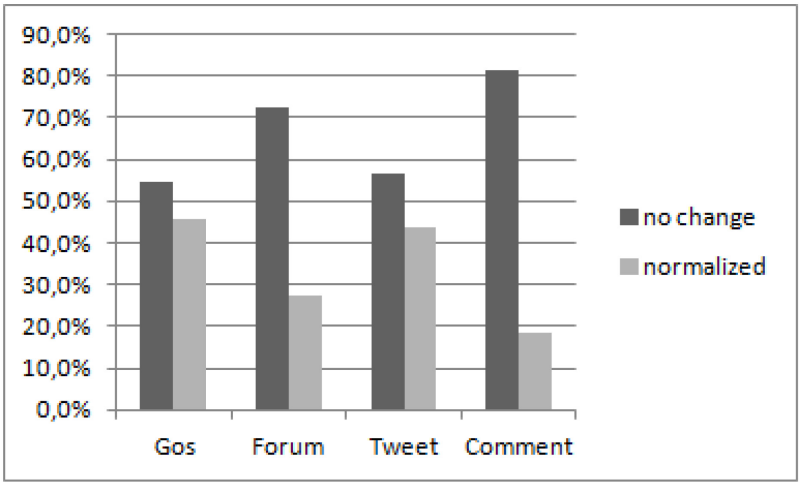

Figure 2: Degree of standardization changes needed in the Gos and Janes corpora.

The results show that in the Gos corpus, a little more than a half of the keywords (55%) were normalized. The normalization is mostly related to pronunciation variation because of reduction on most common words (adverb (44%) and verb (39%)).

Example 8)

Example	**in drgač ne prideš gor k je tok strmo**
Normalization	in drugače ne prideš gor ker je tako strmo
Translation	*and otherwise you won't get there because it's so steep*

As can be seen from Example 8, the most common phenomenon of pronunciation variation in the corpus of spoken Slovene is non-stressed vowel reduction. Besides this phenomenon, pronunciation variation concerns different phonetic levels (neutralization, monophthongization, diphthongization) varying from one dialect to another. Some informal words have gone through numerous phonetic changes and have a very different form compared to their standard equicalents (e.g. *pol - potlej, kva - kaj, jst - jaz*). At this point, it has to be mentioned that the results also depend on the transcription conventions of the Gos corpus transcription using the characters of the Slovene orthographic system following as faithfully as possible the realized acoustic forms of words, with the principal aim to show the typical deviations to the standard pronounciation, see Verdonik et al. (2013).

Regarding the Janes subcorpora, the need for standardization is mostly due to non-canonic spelling (e.g. *drgač/drugače - otherwise*) which is influenced by pronunciation variation in spoken discourse, but also the result of omission of diacritics not easily accessed on smartphone keyboards (*mogoce/mogoče - maybe*) and non-standard tokenization (e.g. *nevem/ne vem - I don't know*). A comparison between the Gos and the Janes corpora shows that the degree of normalization needed in Twitter subcorpus (57%) the most resembles spoken discourse.

Example 9)

Example	**haha jst teb to čist resno!**
Normalization	haha jaz tebi to čisto resno!
Translation	*haha I am totally serious!*

As the proportion of words that had to be normalized is higher in the Gos and the Twitter corpora than in the Comments and Forums corpora, we could conclude that spoken and Twitter communication are less standard than that used in Comments and Forums. Yet, as Example 9 shows, the degree of standardization needed is not the only indicator of informal language as communication on Twitter seems to reflect a sociolect of an urban society finding its interactive way to interpersonal communication here and now (as indicated by the frequently used interjection *haha* as an element of reaction to what has been written and the frequent second-person singular pronoun *you* as an indicator of direct interaction).

The Forum and Comments subcorpora show less resemblance with spoken discourse with respect to the degree of standardization required (28% in Forums and only 18% in Comments). It seems that non-canonic language on Forums and Comments is more topic-related: while a patient asking a doctor to explain the results of a medical report will use canonic orthography, but an adolescent discussing his height with his peers will be less devoted to standard language:

Example 10)

Example	**jst sm 17 pa sm vlek 189 -.- a se da kako pomajnšati?**
Normalization	jaz sem 17 pa sem velik 189 -.- a se da kako pomanjšati ?
Translation	*I am 17 and I am 189 cm tall -.- is there a way to get shorter?*

5.3 Categorization

The previous section showed that several dimensions of non-canonic language use cannot be explained by limiting the analysis to the degree of deviation from the norm in a particular corpus as they require a deeper linguistic consideration as well. This is why we performed a categorization process which shows for each of the analysed corpora whether a word belongs to standard vocabulary or to one of the 10 identified categories of non-standard forms. With this process, we wanted to examine the characteristics of user-generated language that are adopted from informal spoken discourse and those that represent innovative elements of written computer-mediated communication.

5.3.1 Canonical elements

The category of standard expressions contains words which did not display any non-canonic characteristics (e.g. *dejansko* - *actually*). The biggest proportion of them is found in the spoken corpus and in the Forum subcorpus. In must be noted, however,

that some of the words could have been classified into other groups with more context analysis (e.g. several standard forms reveal intense interaction with other participants and could have been categorized in the category 'interaction').

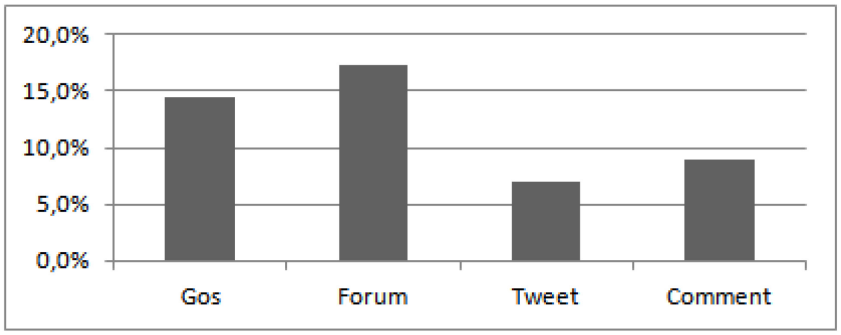

Figure 3: Standard elements in spoken and user-generated corpora.

5.3.2 Spoken language elements

We took a closer look at the non-canonic categories that can be found in spoken and user-generated corpora: non-standard pronunciation or pronunciation-like spelling, topic- or medium-related expressions, discourse markers, and informal or foreign words (Figure 4).

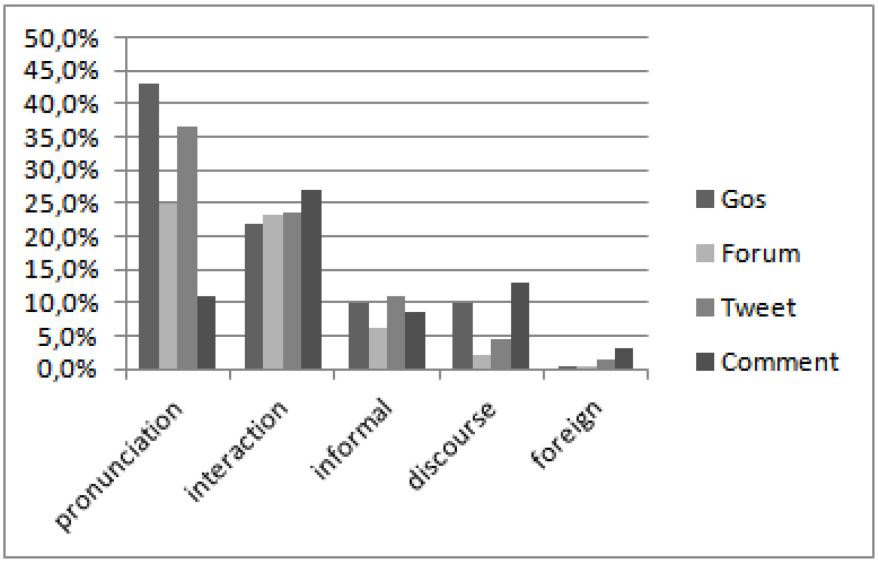

Figure 4: Non-canonic elements present in spoken and user-generated corpora.

Similar to the observations of the standardization process, the Twitter corpus seems to be the most similar to speech in terms of phoneticized spelling of words (43% in Gos vs. 36% in Twitter), interaction (26% in Gos vs. 24% in Twitter), and informal words (10% in Gos vs. 11% in Twitter). As Example 4 shows, the informal words (e.g. *razirat se - to shave, nažajfan - soaped*) co-occur with interaction words (e.g. *sej veš - you*

know) and discourse markers (*jah - well*), which all reflect the relaxed and interactive nature of tweeting:

Example 11)

Example	**jah sej veš.. za razirat se, morš bit nažajfan:)**
Normalization	jah saj veš … ra razirat se moraš biti nažajfan :)
Translation	*well you know … you have to be soaped to get shaved :)*

In the category of discourse markers, the Comments corpus (12%) is the closest to spoken discourse (10%). This category covers mostly adverbs (e.g. *sedaj - now, torej - so*), particles (e.g. *evo - here, pač - well*) and interjections (e.g. *aja - oh, haha*), and gives the impression of imitating the simultaneous process of planning and uttering spoken discourse:

Example 12)

Example	**Haha mi je jasno kako je dobila položaj. Vsaj če držijo besede njenih sodelavcev.**
Translation	*Haha I get it how she got the position. At least if what her colleagues say is true.*

Interactive words are characteristic of all analysed corpora (22–27%) and refer to other participants (e.g. *hvala - thank you*) or to the authors themselves (e.g. *gledam - I am watching*). Deictic expressions (e.g. *tole - this*) and interrogative pronouns, such as *kdo (who)* and *kje (where)* belong to this category as well because they also indicate interaction with other participants.

The biggest outlier in this analysis turns out to be the Forum subcorpus, in which we have detected significantly less pronunciation-like spelling (25%), informal lexemes (6%) and discourse markers (2%) than in the Gos corpus. The degree of use of spoken elements correlates with the degree of formality imposed by the forum topic (e.g. lower in medical discussions, higher in threads on motoring). While Twitter users display a distinctive liking for wordplay and innovative language use, the underlying communicative goal of forum users seems to be much more transactional.

5.3.2 User-generated contents-specific elements

Categories which are only present in the Janes subcorpora but not in the Gos corpus represent the most salient CMC characteristics (Figure 5).

The topic of discussion concerns mostly nouns and is most evident in Forums (e.g. *avto - car, problem*) and in Comments (e.g. *tekma - match, volitve - elections*). We were not surprised by this fact because the Janes corpus was constructed from domain-specific forums and because news comments are by definition topic-specific, unlike the topic-diverse GOS and Twitter data.

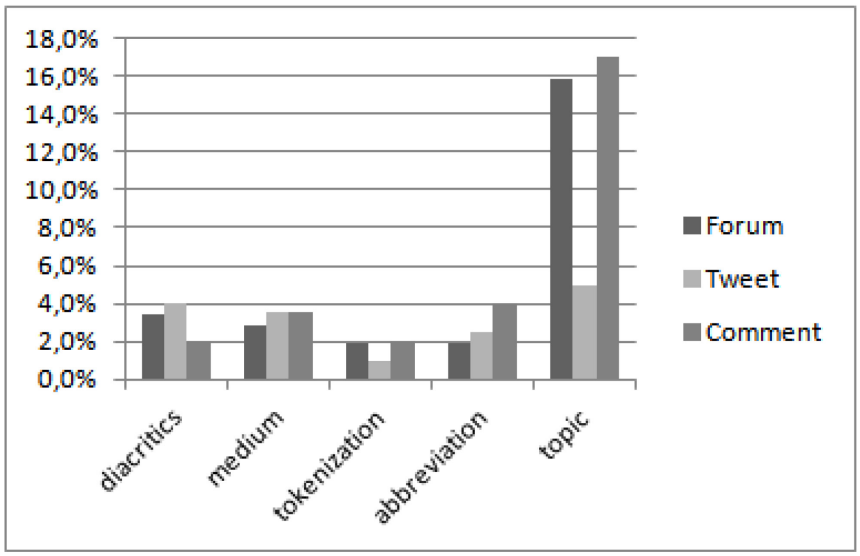

Figure 5: Non-canonic elements only present in user-generated corpora.

All three Janes subcorpora contain keywords revealing the main features of social media (e.g. *com - .com, všeč - like, videoposnetek - video*), the use of which is important because even though they might be limited to a particular medium at first but then become part of the general vocabulary (e.g. *všečkati - like*).

Omission of diacritics, shortening of words and non-standard tokenization are not substantial features in this analysis in quantitative terms because these characteristics are dispersed over different words and will not show within the top typical 200 keywords of a corpus. If a user uses a specific abbreviation, tokenization or does not use diacritic signs, we can only observe the most frequent words characterized by these phenomena. On the level of diacritic signs omission, this is the case of *boš/bos - you will,* while non-standard tokenization also concerns the most frequent verbs (e.g.*ne bi/nebi - I would not*). In our opinion, non-standard tokenization, more often present in Comments and Forums corpora than in the Twitter corpus, reflects the lack of linguistic competence rather than linguistic creativity.

5.4 Linguistic phenomena

In addition to the general non-canonical categories, we tried to identify the specific linguistic phenomenon of each non-canonical keyword. Since more than half of the analysed words did not get a linguistic label because the phenomenon was already sufficiently defined within the categorization process (discourse marker, interactive words etc.), this subcategorization only relates to some categories of the non-standard analysed words (phonetic spelling, informal and foreign words and discourse markers), which is why the results in Figure 6 are accordingly lower.

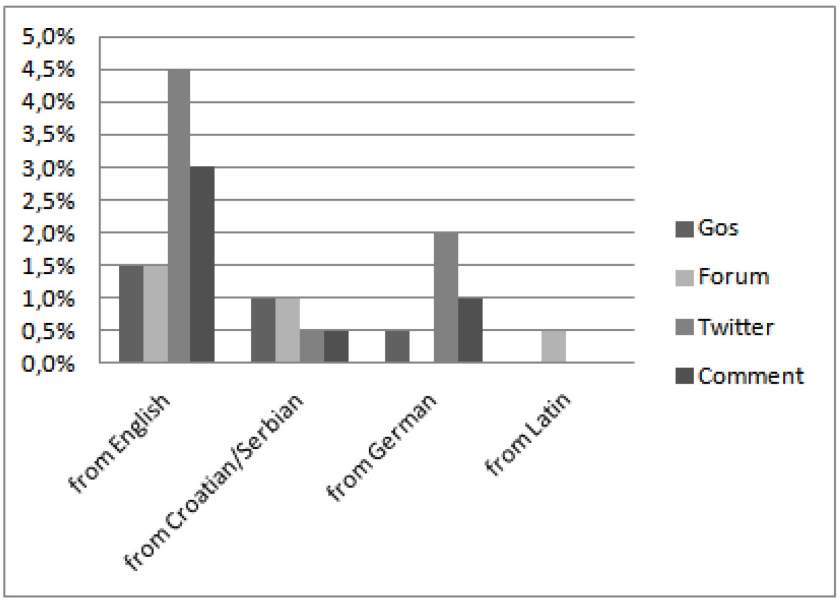

Figure 6: Pronunciation-related phenomena in the spoken and user-generated corpora.

Within the categories that were analyzed in the Gos corpus, the most frequent linguistic phenomena are phonetic reduction, posteriorization, and neutralization, which is also the case for Twitter and Forums (e.g. *drgač/drugače - otherwise*). In order to prevent premature speculation about the nature of pronunciation and spelling tendencies in contemporary Slovene, a larger amount of spoken and user-generated data should be studied.

Foreign words in Slovene have historically been subject to numerous stereotypes and different linguistic perspectives have shown very diverse attitudes. As Figure 7 shows, elements from four languages were identified among the top 200 analysed keywords. In the corpus of spoken Slovene, three words were derived from English (*jes – yes*), one from Croatian or Serbian (*kao - like*) and one from German (*fajn - fein*[2]). Among the user-generated corpora, the Twitter and the Comment corpus seem to contain the most foreign words, considerably more than the analyzed spoken data. On Twitter, we found seven words derived from English (e.g. *app, top*) and four from German (e.g. *direkt, ziher*), while within Comments, six words were from English and four from German. As we do not want to jump to any premature conclusions with respect to the status and trends of foreign word usage in user-generated contents, a more thorough analysis is reserved for future work.

[2] This expression could also have been classified as an English one, but due to the historic influence of German in Slovene, we categorized it as a German word.

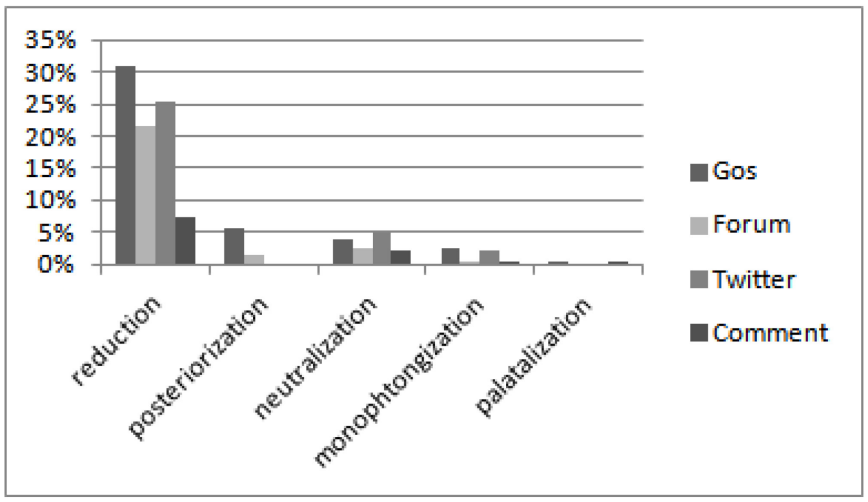

Figure 7: Foreign words in the spoken and user-generated corpora.

Other interesting linguistic phenomena that we have detected are the frequent use of deixis (*tale - this one, tam - over there*), typical in spoken discourse but also characteristic of user-generated corpora, and the presence of "articles" which do not exist in traditional Slovene language manuals (*una ta vesela - the happy one*).

6. Discussion of the results

The qualitative and quantitative analysis performed in this study expose the most salient phenomena that show common points and discrepancies between the compared corpora. The first column of Table 6 (*Spoken language*) presents the typical features of spoken discourse compared to the written standard Slovene; the second column (*Similarities*) displays the user-generated subcorpora that contain most of the detected spoken elements; and the third column (*Differences*) relates to the detected specifics of user-generated corpora that are not present in the spoken corpus.

	Spoken language	Similarities (example)	Differences (example)
normalization	high level (45%)	Twitter (*jst - I*)	Comments; standard words (*politiki - politicians*)
categorization	pronunciation (43%)	Twitter (*drgač - otherwise*)	Forums; topic-related vocabulary (*original - original*)
	interaction (21%)	all corpora (*strinjam - I agree*)	
	informal (10%)	Twitter (*ziher - for sure*)	Comments; topic-related vocabulary (*krivi - guilty*)
linguistic phenomenon	reduction (31%)	Twitter (*dobr - well*)	Comments; 1 instead of 2 words (*nebi - wouldn't*)
	deixis (4%)	Forums; deixis (ta - *this*)	
	foreign words (2%)		

Table 6: Similarities and differences between spoken and user-generated corpora.

The results show that the spoken and the Forum corpora have similar proportions of adverbs and verbs, but that the Twitter corpus shows the most similarities with spoken discourse on the levels of non-standard pronunciation and spelling variants, interaction words and informal lexemes. The most salient specific characteristics of the Comments corpus are a higher proportion of nouns than in speech and a lower level of normalization required compared to speech, while in Forums, topic-related words and non-standard tokenization are prolific.

7. Conclusions and future work

This paper presents a language-independent triangular methodology for lexical comparison of the entire spoken–written spectrum with user-generated content and its informal communication falling roughly in the middle. The results show that a considerable amount of various spoken-language characteristics permeate computer-mediated communication. This is why these characteristics are gaining in importance as they are acquiring new functions in the increasingly interactive and instantaneous online communication where the line between spoken and written discourse are blurred. For this reason, the treatment of such phenomena in contemporary lexicography needs to be re-examined and updated.

It must be noted, however, that this is only the beginning of our studies on this topic which will be extended beyond lexical level in our future work in order to comprehensively also include the context of words (i.e. phraseology, collocations, colligations, syntactic structure). We expect the greatest need for methodological changes at the syntactic level where traditional approaches via conjunction analysis cannot be used and a more important focus should be given on text comprehensibility. Regarding the detected particularities of user-generated communication, a more focused analysis should be carried out on omission of diacritics, word-shortening strategies and non-canonical tokenization.

8. Acknowledgement

The work described in this paper was funded by the Slovenian Research Agency within the national basic research project “Resources, Tools and Methods for the Research of Nonstandard Internet Slovene” (J6-6842, 2014-2017).

9. References

Akinnaso, F. (1982). On The Differences Between Spoken and Written Language. *Language and Speech,* 25/2, pp. 97–125.

Atkins, S.B.T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.

Bamman, D., Eisenstein, J. & Schnoebelen T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics,* 18/2, pp. 135–160.

Baron, N. (2010). Discourse Structures in Instant Messaging: The Case of Utterance Breaks**.** *Language@Internet 7*, article 4.

Beißwenger, M. (2013). DeRiK: A German reference corpus of computer-mediated communication. *Literary and Linguistic Computing,* 28(4): pp. 531–537.

Chovanec, J. (2009). Simulation of spoken interaction in written online media text. *Brno Studies in English,* 35/2, pp. 109–128.

Crystal, D. (2007). *How language works.* New York: Penguin Books.

Fišer, D., Erjavec, T., Zwitter Vitez, A. & Ljubešić, N. (2014). Janes se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino. In T. Erjavec & J. Žganec Gros (eds). *Language technologies: proceedings of the 17th International Multiconference Information Society - IS 2014,* pp. 56–61.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. *Proceedings EURALEX 2004.* Lorient, pp. 105–116.

Kosem, I. (2015). Fran, pameten in intuitiven ? *Slovenščina 2.0/2*, pp. 161–193.

Leech, G., Rayson P. & Wilson A. (2001). *Word Frequencies in Written and Spoken English: Based on the British National Corpus.* London: Longman.

Logar Berginc, N. & Krek S. (2012). New Slovene Corpora within the Communication in Slovene Project, *Prace Filologiczne,* 63, pp. 197–207.

Ljubešić, N., Erjavec T. & Fišer, D. (2014). Standardizing tweets with character-level machine translation. In A. Gelbukh (ed.) *Computational linguistics and intelligent text processing : 15th International Conference.* Heidelberg: Springer, pp. 164–175.

Morel, M.A. & Danon Boileau, L. (1998). *Grammaire de l'intonation.* Paris: Ophrys.

Moon, R. (1998). On using spoken data in corpus lexicography. In T. Fontenelle, P. Hiligsmann, A. Michiel, A. Moulin, S. Thiessen (eds.) *Euralex 98 proceedings.* Liège: University of Liège, pp. 357–362.

Pavesi C. (2014). Features of Speech in a Corpus of Learner English CMC: the case of "a lot of" In A.C. Murphy & M. Ulrych (eds.) *Perspectives on Spoken Discourse.* pp. 61–79.

Rundell, M. (2014). Macmillan English Dictionary: The End of Print? *Slovenščina 2.0, 2*, pp. 1–14.

Sinclair, J. (1987). *Collins Cobuild English Language Dictionary.* Collins.

Sindoni, M.G. (2013). *Spoken and Written Discourse in Online Interactions: A Multimodal Approach.* New York/London: Routledge.

Verdonik, D. & Zwitter Vitez, A. (2011). *Slovenski govorni korpus Gos.* Ljubljana: Trojina, zavod za uporabno slovenistiko.

Verovnik T. (2004). Norma knjižne slovenščine med kodifikacijo in jezikovno rabo v obdobju 1950–2001. *Družboslovne razprave XX,* 46/47: pp. 241–258.