The expansion of the Gigafida corpus: Internet content

Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Nataša Logar and Vesna Mikolič

Abstract

The paper discusses the expansion of the Gigafida corpus, a Slovenian reference corpus, to include Internet content, i.e. web pages and user-generated content (tweets, blogs, forums and comments on news portals). The resources and tools available which are best suited to achieve this objective are discussed, and the web crawling methodology used for this purpose is also presented.

Keywords: reference corpus, Slovenian, dictionary, Internet content, web crawling

1 INTRODUCTION

In Logar Berginc et al. (2012: 45) we opened the chapter entitled "Web Text in the Gigafida Corpus" with the finding that the written language is becoming less commonly used in the form of the printed word, and more commonly seen in electronic media. The chapter presented data showing that as of October 2007, 66% of the respondents, aged between 12 to 65 years, were using the Internet (RIS survey).[1] The most recent percentages are – as expected – even higher: according to an analysis by the Statistical Office of the Republic of Slovenia, in the first quarter of 2014, 97% of Slovenian households with children and 70% of households without children had Internet access, and during this time 72% of all people aged 16 to 74[2] years old were using the Internet. It may be added that

> 81% of these persons /.../ were using the Internet every day or almost every day. The largest percentage (87%) used it for sending or receiving e-mails and finding information about goods or services. / 58% of respondents in the first quarter of 2014 participated in online social networks (in the first quarter of 2013 the figure was 53%) (ibid.).

Another important finding was that 66% of the users accessed the Internet via mobile phones or other mobile devices (e.g. a tablet). The Internet is thus accessible anywhere, and not just for reading, watching and listening, but also for writing and publishing texts, images, music, and so on. Widely available public platforms that rely on language – once limited to print, radio and television – are now open to contributions from virtually everyone, and this has brought a new kind of Slovenian into public use: texts showing linguistic characteristics that were previously primarily used for speech in private and informal situations.

Editors of modern reference corpora of different languages include web texts into their work in various different ways. The overview presented by Logar Berginc and Ljubešić (2013) noted "a common tendency for including texts from the Internet in the reference corpus, although to what extent this may happen in the future is not yet clearly defined, but if the corpus already contains or will contain texts from the Internet, texts of different genres should also be included" (ibid: 103). Consequently, on the one hand we have for example the *Oxford English Corpus*, from which *Oxford Dictionaries* arise,[3] that is almost entirely composed of texts from the Internet, and on the other hand, for example *The Slovak National Corpus*, on the basis of which *Dictionary of Contemporary*

1 http://www.ris.org/
2 http://www.stat.si/StatWeb/glavnanavigacija/podatki/prikazistaronovico?IdNovice=6560
3 http://www.oxforddictionaries.com/

Slovak Language,[4] is currently being prepared, that does not contain any online texts (see more in Table 1 in the chapter Reference Corpora Revisited: Expansion of the Gigafida Corpus).

As will be seen later in this chapter, we see texts from web pages, comments on news sites, blogs, tweets and forum messages as a significant part of the public written Slovenian, which is why we argue they should be included in the corpus that will be the basis for the future reference dictionary of our language. As such, lexicographers should be interested in lexicons that are used in different circumstances by all the speakers of Slovenian, not just journalists, translators, writers, and so on. We should therefore pay special attention to (semi)public written online communication that is determined by circumstances such as (non)interactivity (a)synchronicity, physical (non)presence/absence of the interlocutor and other situational factors, resulting in a highly interactive form of communication with more elements of the spontaneous spoken language, and with (adapted for computer communication) paralinguistic and prosodic elements (Crystal 2001). The task of a corpus as a lexical resource *must* therefore be also to capture this linguistic reality, so in this chapter we illuminate this issue from four angles:

a) the initial state of the Gigafida corpus (compared with the slWaC$_2$ corpus, the online corpus of Slovenian),

b) diversity of online text genres and reasons for their inclusion in the corpus (or exclusion from it),

c) resources and tools that are already available for a future upgrade of the Gigafida corpus (the JANES project),[5] and

d) the most appropriate methodology of web crawling, including the possibility of building a subcorpus that would be regularly updated.

2 GIGAFIDA AND SLWAC$_2$: EXISTING STATE, COMPARISON, BINDING POSSIBILITIES

Web sites that were included in the Gigafida corpus and technologies for their collection are described in more detail in the already mentioned chapter in Logar Berginc et al. (2012: 45–67), so we shall only note that integrating web content into the Gigafida corpus was "methodologically speaking, the first such major attempt in Slovenia that could formulate guidelines for the future construction of Slovenian reference corpora and indicate some interesting comparative linguistic

4 http://slovniky.juls.savba.sk/
5 http://nl.ijs.si/janes/

analysis" (ibid.: 45). Gigafida therefore contains texts from 10 news portals and a total of 91 introductory web pages (29 corporate web pages, and 42 cultural, state, research and university institution web pages). The web was crawled in the period April 2010 – April 2011, and it contributed more than 185 million words to the corpus, of which 63% come from news portals (24ur.com, rtvslo.si, siol.net, etc.), 30% from institutional web pages (gov.si, uni-lj.si, sazu.si, ijs.si, etc.), and 7% from corporate web pages (eles.si, gorenje.si, and kolosej.si, among others). The procedure to capture texts from web pages followed several steps: selection and preparation of the programme for three regimes of crawling (daily, monthly and one-off), boilerplate removal, language detection, and finally the detection and the removal of duplicates and near-duplicates. It turns out that in order to achieve seemingly simple tasks, i.e. that of including web texts in the reference corpus, a fairly complex methodology is required, that – along with the criteria of selecting of web sites and rating of the obtained results – we successfully tested and adapted for use with Slovenian (more on the latest methods of crawling are reported in section 5, below).

During the integration of web texts in the Gigafida corpus – in 2011 – a new and methodologically similar corpus of Slovenian emerged, the corpus slWaC (Erjavec and Ljubešić 2011),[6] which was upgraded to $slWaC_2$ in 2014 (Erjavec and Ljubešić 2014). $slWaC_2$ contains 1.2 billion words from texts acquired from over 37,000 web domains or 2.8 million URLs. The methodology of the construction of the two versions of slWaC is presented in detail in the references mentioned and in Logar Berginc and Ljubešić (2013: 87–89).

The existence of two large corpora of Slovenian has prompted some comparisons that have shown what both of them contain, as well as what their deficiencies are (as much as a comparison of the two entities can reveal in this regard). A comparison based on the frequency profiles (Rayson and Garside 2000) of Gigafida and $slWaC_2$ showed (Erjavec et al. 2015b: 40) that in the latter there are several texts related to computer science, the Internet and the use of web contents, while Gigafida contains more texts that are typical for newspapers, on subjects such as sports, domestic politics, the economy and crime.

To the already published comparative data (ibid., and in Logar Berginc and Ljubešić 2013), we now add data from more recent comparisons between Gigafida and $slWaC_2$, obtained by the topic modelling method (Blei et al. 2003; Sharoff 2010) – though here only paying attention to possible weakness of the Gigafida corpus. Tables 1 and 2 show the 20 most common topics for Gigafida and $slWaC_2$, respectively.

6 http://nl.ijs.si/

Table 1: Noun lemmas, which most likely belong to one topic, and the occurrence of the topics in Gigafida.

Subject	Frequency*	Noun lemma
people, family, life in general	4,835	otrok leto dan čas ženska življenje človek družina oče moški roka prijatelj glava žena mama mož sin starš hiša
sport	4,034	tekma mesto leto ekipa zmaga točka igra sezona igralec prvenstvo klub liga prvak trener minuta konec pokal krog reprezentanca
domestic politics	3,639	predsednik vlada država stranka svet minister leto zakon volitev predlog poslanec vprašanje komisija član odbor zbor seja politika ministrstvo
society, OTHER	3,631	človek življenje svet čas odnos način stvar država vprašanje družba primer beseda delo moč stran problem resnica leto občutek
shows, performances etc. in the local area	2,865	ura društvo leto prireditev dan sobota dom član vas mesto občina skupina nedelja šola srečanje gost obiskovalec dvorana delo
war, terrorism, criminal acts	2,669	leto vojna država policija policist človek vojska dejanje orožje dan napad vojak žrtev sodišče oblast zapor kazen čas mesto
TV and radio programmes	2,622	film leto glasba oddaja tv poročilo skupina serija dan pesem festival čas koncert predstava program vloga gledališče del novica
traffic	2,481	cesta pot dan nesreča leto ura voda voznik morje vozilo mesto meter promet letalo kilometer čas vožnja kraj avtomobil
economy	2,470	leto odstotek država podjetje cena trg plača izdelek rast razvoj delo gospodarstvo proizvodnja delavec število področje strošek mesec sistem
education	2,363	šola delo leto otrok program področje znanje študent projekt izobraževanje univerza fakulteta učenec razvoj starš organizacija učitelj center zavod
finances	2,292	milijon evro tolar leto banka družba podjetje odstotek delnica milijarda dolar denar vrednost cena prodaja delež dobiček trg sklad
local politics	2,228	občina leto prostor gradnja objekt cesta projekt območje delo zemljišče mesto milijon stanovanje okolje podjetje načrt denar voda tolar
animals, nature, living spaces	2,153	žival barva prostor vrsta pes voda hiša gozd del material les vrt tla drevo konj čas leto vrata oblika
law	2,145	zakon člen sodišče postopek pravica primer podatek organ odstavek dan oseba podlaga pogodba delo odločba sklad stranka določba zadeva

Subject	Frequency*	Noun lemma
publications, culture, art	2,087	leto knjiga delo razstava stoletje cerkev mesto čas muzej svet ime zbirka umetnost avtor jezik zgodovina del slika beseda
motoring	2,021	m sit avtomobil motor km cena vozilo eur d e l model leto avto x n g r h
health	1,942	bolezen zdravnik bolnik zdravilo telo človek zdravljenje leto koža težava dan zdravje primer rak bolnišnica bolečina kri celica čas
media	1,788	naslov stran številka medij novinar revija dan nagrada pošta časopis leto ime delo informacija oddaja članek televizija bralec vprašanje
information and communication technology	1,491	računalnik sistem uporabnik podatek program slika stran naprava uporaba kartica telefon zaslon internet omrežje model oprema tehnologija možnost storitev
food	1,437	vino voda olje rastlina minuta meso sladkor g sol hrana zelenjava jed žlica okus sadje mleko krompir sok list

* "Frequency" in the second column signifies the occurrence of individual topics in the corpus.

Table 2: Noun lemmas, which most likely belong to one topic, and the occurrence of the topic in slWaC$_2$.

Subject	Frequency	Noun lemma
people, family, life in general	3,929	otrok dan čas leto človek ženska roka pes življenje stvar prijatelj moški glava mama ura družina starš svet konec
society, OTHER	3,266	človek življenje svet čas način odnos stvar družba otrok ljubezen beseda primer vprašanje resnica pot občutek ženska moč problem
domestic politics	2,626	vlada država predsednik stranka zakon svet leto predlog minister član poslanec komisija vprašanje zbor odbor politika skupina pravica mnenje
travelling, tourism	2,524	pot mesto dan cesta ura leto čas vrh morje voda smer gora meter del dolina kraj gozd hotel stran
economy, development	2,360	podjetje področje razvoj sistem projekt delo leto trg storitev okolje država cilj organizacija program izdelek znanje rešitev sodelovanje tehnologija
finances	2,265	leto evro odstotek milijon podjetje banka država cena družba denar trg vrednost rast milijarda sredstvo delnica plača prodaja mesec
sport	2,232	tekma ekipa mesto igra leto točka zmaga sezona igralec minuta prvenstvo klub liga konec tekmovanje prvak rezultat trener pokal
shows (film, music, theatre)	2,139	film leto glasba skupina album pesem festival koncert skladba čas oder nastop predstava nagrada vloga dan zasedba oddaja zgodba

Subject	Frequency	Noun lemma
education	2,072	šola otrok leto delo program študent učenec znanje starš ura izobraževanje fakulteta univerza čas študij delavnica področje učitelj dan
health	2,059	telo bolezen koža zdravilo težava zdravljenje zdravnik dan leto bolnik bolečina človek zdravje celica primer čas kri otrok učinek
online shopping	2,042	stran podatek uporabnik naslov storitev vsebina račun pošta cena ime nakup internet številka informacija izdelek naročilo ponudba dan paket
law	2,016	člen zakon sodišče postopek pravica odstavek oseba pogodba primer dan stranka podlaga sklad organ delo določba odločba podatek pogoj
local politics	1,937	občina leto projekt društvo območje mesto delo prostor sredstvo objekt program član center gradnja zavod organizacija področje okolje ministrstvo
religion	1,887	leto cerkev človek vojna bog življenje dan mesto čas smrt vojska svet država oče maša ime beseda vera stoletje
publications, culture, art	1,826	leto knjiga delo jezik razstava avtor medij beseda fotografija nagrada zbirka revija del umetnost zgodba naslov čas svet dogodek
information and communication technology	1,781	računalnik naprava sistem slika program telefon fotografija podatek uporabnik uporaba video zaslon stran aplikacija dokument model kamera različica oprema
motoring	1,697	vozilo avtomobil motor barva vožnja voznik model kolo avto del leto oblačilo znamka cesta hitrost obleka sedež oprema sistem
living spaces	1,624	voda prostor energija hiša material sistem površina odpadek zrak objekt naprava del uporaba stanovanje temperatura okno okolje les plin
food	1,471	hrana voda olje rastlina vino mleko okus meso zelenjava vrsta jed sadje izdelek oseba količina dan kislina žival sladkor
World Wide Web	0,520	piškotek dan nastavitev seja mesto namen stran storitev uporaba informacija podatek oglaševanje klik gumb primer ura facebook možnost novica

Three topics can be identified that are of particular importance when selecting URLs to obtain new web texts to upgrade the Gigafida corpus (tweets, forum messages, comments on news sites and blogs are discussed in the next section). These are topics that the current Gigafida corpus, with mostly printed texts and only a small and narrowly selected set of web texts, has poor coverage of, and

these thus needed to be examined if we want to describe their distinctive lexicons in a dictionary. For slWaC$_2$ (but not for Gigafida) the typical topics are *travel, tourism, online shopping* and the *World Wide Web* (see last row in Table 2). The topic *religion* is in this respect surprising, because it is the only one that could be better integrated into Gigafida by means of printed texts (wherein the response of the text providers is crucial).

At the end of such comparisons the question of the direct inclusion of the web corpus of Slovenian, slWaC$_2$, into the new Gigafida corpus arises. From the perspective of a more focused and controlled, as well as time-predictable and equable, form of text collection, with the explicit purpose of inclusion in the reference corpus, this question would be better answered in the negative, but it is not necessary to keep future upgrades of both corpora completely separate. On the contrary: as will be shown in section 5, these corpora are closely connected by their method of construction. Furthermore, the existence of two corpora of modern Slovene is also useful in terms of synergies, and as a demonstration of their differences and deficiencies.

3 WEB TEXT GENRES AND DICTIONARY SOURCES

On the Internet, the most influential medium of the 21st century, we are faced with a variety of communication environments or areas that apply all four basic functions of text (Skubic 1995; Mikolič 2007): cognitive, communicative, executive and art-expressive. There are also various discourse/speech communities that determine the characteristic language choices people make in the context of a specific discourse/speech.

Some of the features of web texts are tied to (more) informal speech situations, these are often manifested in texts in non-standard form (e.g. slang, jargon, vernacular language, and dialect). On the other hand, other web texts correspond to the concept of public communication in the narrow sense of the word (Škiljan 1999), and are written in accordance with standard language norms. The language heterogeneity of the Internet has caused changes in the language and the expansion of its lexis, so it is necessarily to find out which texts must be an integral part of any corpus that will be the basic source for a dictionary of modern Slovene (and at the same time, we can find out which texts it is possible, at the moment, to reject).

A description of the variety of online genres and their key factors is actually a rather difficult task, due to the extensiveness and uncontrollability of the material,

and the small number of studies of such genres and their target audience (Crowston 2010: 17, 26).

Nevertheless, on the basis of the analysed literature and related material, it can be seen that, for the analysis of online text genre variety and also for establishing the selection of web texts for the corpus, there are two key criteria, as presented by Herring et al. (2004):

- authorship or the relationship between the sender and the recipient (one or more authors, a formal or informal relationship) (see also Oblak et al. 2005),
- functions and the associated internal and external structure or form of the text, as well as multi-codes and updates (see also Bishop 2009; Crowston 2010).

The language choices of online authors depend on both these criteria, particularly in relation to conformity with the norms of standard language, or deviations from them.

From the perspective of the author, web texts are basically divided into:

- classic websites (HTML) with one single author or source of the texts,
- online community genres ("web-based community genres", Bishop 2009) with more than one author of the texts,
- blogs/blog writings (blogging) as an intermediate genre between one- and two-way communication.

a) Classic websites are mainly characterized by one-way communication (the most common exception here are media sites, which may include the forum messages, comments, or blog writing of the readers). The source of a website's text source is known or easily determinable. The relationship between sender and the recipient is mostly formal, and since texts address the general public they are mostly written in accordance with the standard language norms.

Among classic websites we place the following:

- Web portals (as well as Wikipedia, Wikisource, Wikiversity, Wikibook, and so on),
- media sites,
- commercial and corporate websites,
- websites of governmental and non-governmental organizations and local government bodies.

Personal websites are less formal and may be closer to the form and the purpose of blogging or the genres of online communities (such as on Facebook).

b) Online community genres ("web-based community genres") are related to collective-action oriented websites or interactive text forms of computer mediated communication (CMC), in which several authors collaborate. These genres are determined by the dominant actors, communication environment or topic, and the internal structure. The language choice here also determines the nature of the interactions among the actors. They are very diverse and often also anonymous, so the expressiveness of the texts is rather varied and they mostly include elements of vernacular, informal language genres. These genres are increasingly replacing speech communication, and so they are often manifested by written spoken language, and in some applications also by spoken text.

Among the online community genres we could place texts from various online tools and social networks, such as:

- forum messages (users are discussing a certain topic)
- Twitter, Facebook, Myspace, and LinkedIn (texts such as tweets, statuses or thoughts, status comments, photos, videos, hyperlinks, interest groups, event creation, invitations etc.),
- Instagram (publishing photos and hyperlinks to Twitter or Facebook),
- Ask.fm (users create an account and other users then ask them questions, also using hyperlinks to Twitter or Facebook),
- Snapchat (a mobile application through which users share thoughts, videos, photos etc. with their friends. Their messages disappear in a few minutes),
- Viber (a mobile application for smartphones, through which communication takes place via the Internet, can be written or spoken, and includes mobile contact list),
- web chats (diverse categories of "rooms" where users with same interests connect with each other).
- Comments on journalistic articles, videos and so on (comments can then develop into a discussion on a specific topic, usually between users, unknown to each other, and this functions according to the principles of a forum).

c) Blogs/blog writings are most often part of the journalistic genre, intended for a wider audience, and are often also in direct interaction with the readers. Usually there is one single author of a blog, and these can be written by professional (journalists) or unprofessional writers, so the language choice depends on the

communication competence of the author and especially on the target audience that the author wants to reach.

According to Domingo and Heinonen (2008) we can distinguish the following types of journalistic blogs/blog writings, which are differentiated by professionalism of the writers and degree of institutionalisation of the environment:

- citizen blogs (written by unprofessional writers outside media institutions),
- audience blogs (written by unprofessional writers within media institutions),
- journalist blogs (written by journalists outside media institutions),
- media blogs (written by journalists within media institutions).

In terms of *function*, texts are classified in groups of broad text genres and narrow text types and according to the following common properties: the purpose or influential role, recipient, reference and external and internal structure of the text (comp. Mikolič 2013; Nidorfer Šiškovič 2013). According to these properties, we also analysed the text in a web environment, where we referred to Crowston (2010) who summarizes the key typologies of Internet genres considering the purpose and form. Based on the findings of this, we can try to describe web genres in the context of the following groups (as summarized by Mikolič and Rolih 2015):

1. *Conversational and at least partially private text genres*: e-mail, web chats, tweets and other genres of social networks (e.g. Twitter, Facebook) and forum messages. The adjective "private": for these genres is based on their greater range of private language elements or content components of private communication spheres (Škiljan 1999), sociolects and idiolects (Skubic 2004).
2. *Promotional, advertisement and commercial text genres:* banner adverts, link collections, online shops, marketing and sales websites, personal websites, often with the purpose of self-promotion and marketing. The aim of these genres is to influence the consumer behaviour of the recipients. Due to their appeal to the general public, the language used in these genres generally does not depart from standard language norms, except when stylistic effects need to be achieved.
3. *Reporting/news and broadcast journalistic text genres:* journalistic texts of various genres, the online editions of print media, contributions associated with lifestyle (e.g. recipes, tips in the form of tutorials, guides for a healthy body, and so on). These are genres in which deviations from the standard language norms have only the stylistic role. The exceptions are the comments below the contributions on the major news portals, in which the authors do not usually follow such norms.

4. *Program text genres:* technical data/assistance/support, problem reports, and frequently asked questions (FAQ). These texts are messages from the operators or programmers of web pages. The text opens by discussing a problem and leads the user to a solution. Since it is a professional text aimed at the general public, the language is mostly consistent with standard language norms.
5. *Academic text genres* (accessible at sites such as Google Scholar): technical and scientific texts, written in line with standard language norms.
6. *Official and officiated text genres*: records of the meetings of state bodies, legislative websites, stock market websites, published policies, and so on; online administration, e-applications, etc. The purpose of these genres is to inform the general public about the key procedures, rules and laws in the country, and to enable working with the administrative authorities through the use of online forms. The language in these texts thus does not deviate from the standard language norms.
7. *Literary and semi-literary text genres*: these are belletristic texts, which are characterized by compliance with the standard language norms with an intentional deviation from it. The most common semi-literary web text genres are blogs and web diaries.

Undoubtedly, most online text genres – although still under-explored in both Slovenia and internationally – are very active in terms of their implementation and the development of language.

Due to the rapid development of online tools, some web genres may quickly become out-of-date (at this moment, for example, we are seeing the decline of web chats) and others will emerge, with similar or completely different intentions and linguistic characteristics. Therefore, in the preparation of the dictionary descriptions, we should not only *consider* the online linguistic reality, but also regularly *follow* it.

Of the various web text types described above, the new version of Gigafida should at least consider the content that has a known author or source and it intended for the general public. These texts should include those from large, mainstream websites and personal websites with large readerships, professional writers' blogs, the tweets and Facebook pages of individuals and institutions that have a great impact on general linguistic use (based on number of followers and media responses). Therefore, in terms of function, these texts include some of the conversational, promotional, advertising and commercial text genres, and all of the reporting/news and broadcasting, official and officiated and literary and semi-literary web text genres. As mentioned before, the main conditions for conclusion must be a high level of influence and large readership, and that the text's genre should be evident from the taxonomic categories.

4 USER-GENERATED CONTENT

A special challenge in contemporary lexicography is the vocabulary in user-generated content, published by regular people, and not professional writers. This kind of computer-mediated communication (CMC) is heavily characterized by varying degrees of interactivity, synchronicity and physical detachment. The more the selected medium is interactive, the more elements of spoken language it displays, including the CMC-adapted paralinguistic and prosodic elements (Crystal 2001). The most common features of this kind of language are non-canonical spellings, colloquial and regional expressions, foreign-language elements, non-institutionalised abbreviations, as well as neologisms. These make such texts extremely valuable for lexicographic purposes, but they are at the same time very difficult for automatic processing (Sproat idr. 2001), which is why the development of tools that can handle noisy texts from the web is currently one of the most active research topics in natural-language processing.

In contemporary linguistics, paradigms that consider non-standard language variants in computer-mediated communication as a sign of imperfect or impoverished communication abilities have become a thing of the past, since a number of studies have demonstrated that users adapt their language to maximise the potential and the functionalities of the medium in order to meet their communication needs with the least time and effort required, displaying their identity and spontaneous speech along the way (Herring 2001).

The discrepancy between the language as a living organism, and its static description that calls for research into non-standard language, has been addressed by several Slovenian linguists who have analysed the language of text messages, forum posts as well e-mail (cf. Kalin Golob 2008; Jakop 2008; Michelizza 2008). However, this kind of research is still not receiving enough attention by the mainstream linguistic community, and, as a consequence, the Slovenian linguistic landscape lacks a comprehensive description of the non-standard language varieties, as well as sufficient, publicly available collections of such text types.

JANES, the basic national research project, aims to close this gap and develop the resources, tools and methods need for the analysis of CMC (Fišer et al. 2014a). This section presents the interim results of the projects relevant for the construction of a modern dictionary of Slovene.

4.1 The JANES corpus of Slovenian user-generated content

The current version of the JANES corpus contains four types of user-generated content: tweets, forum messages, news comments and blog posts. Tweets have been

harvested for the past two years with a custom-built tool called TweetCat (Ljubešić et al. 2014). One-off crawling of forum messages and news comments was performed using designated crawlers and text extractors of some of the most popular or influential forums and news portals, based on their traditions, forms of text production and the number of users. Blogs were adopted from the de-duplicated version of the slWaC 2.0 corpus (Erjavec and Ljubešić 2014) by using the string "blog" in the domain name as a positive filter. This is only a temporary solution, as the lack of an internal structure of blogs makes it difficult to distinguish between the language of the main text of the blog and the language of the readers' comments on it. A designated crawler and text extractor for blogs that takes this into account will therefore be developed for the next version of the corpus.

All the texts along with the unified metadata are merged into the JANES corpus and formatted in a bespoke XML, thus enabling corpus structuring, metadata labelling and Unicode character encoding. The corpus is also annotated. Sentence segmentation and tokenization was performed with the standard mlToken library for Slovenian which is part of the ToTaLe (Erjavec et al. 2005) tool chain. Next, word forms were normalized with a character-based machine translation approach that was trained on 1,000 manually normalized key words obtained from the tweet corpus with respect to the reference corpus KRES (Ljubešić et al. 2014). Finally, the corpus was morphosyntactically tagged and lemmatized with ToTaLe, which was originally developed for standard Slovenian.

The JANES v0.3 corpus comprises 161 million tokens, most of which come from tweets (38%), followed by forum messages (29%), blog posts (24%) and news comments (9%). The corpus is already a useful resource for lexicographic work, since it is complementary to the reference Gigafida corpus in terms of content, is substantial in terms of size, and diverse in terms of the text types included. Further enhancement of the corpus by increasing the number of text sources, especially forums and news comments, would of course be highly desirable. It also needs to be noted that while the JANES corpus is limited to public CMC, lexicographers would benefit greatly from the private communication on social media, such as Facebook, which has 750,000 Slovenian users, as well as the new apps that are becoming popular with younger users, such as Instagram and WhatsApp, but also multimedia and video technologies, such as YouTube, Skype and FaceTime, that are taking the place of the traditional text messaging and on-line chats, as seen with MSN Messenger.

4.2 Non-canonical language in the JANES corpus

While it is true that the JANES corpus contains user-generated content, not all of it is written in non-canonical language. Quite the contrary, a quick manual

examination of a small sample of random tweets has shown that a large majority of them are in fact perfectly standard, which may seem surprising at first but since Twitter is used as a popular information dissemination channel, not only by individuals but also by news agencies, public institutions and companies, it is only natural that such communication is carried out in standard Slovenian.

In order to be able to focus on the analysis of non-canonical language, we have developed an approach to automatically measure the level of standardness of the input text at two levels: technical and linguistic (Ljubešić et al. 2015). Technical standardness considers capitalization, use of punctuation and spacing, while linguistic standardness takes into account spelling, lexical choice, word order and so on. A training set of tweets, forum posts and news comments was manually annotated for both standardness levels and scored from 1 to 3, with 1 meaning very standard and 3 very non-standard.

About 30 features that could serve as indicators of technical and linguistic standardness were defined at the character level (e.g. ratio of punctuation written to text length), string level (e.g. ratio of capitalized words written to text length) and word level (e.g. ratio of out-of-vocabulary words written to the Sloleks lexicon). The training set and the features were used to train a linear regressor that assigns a technical and linguistic standardness score to all texts in the corpus, enabling lexicographers to limit their searches to the desired level of standardness.

5 COLLECTING INTERNET CORPORA

Crawling is a process of automatically gathering documents from the web with the purpose of generating search engines indexes, retrieving other information from the web, or building corpora. High recall is the key factor in the former case, while the latter case strives toward acquiring clean linguistic content. Here, it is better to lose parts of the retrieved documents than to get a larger but very noisy corpus, which would contain elements such as the headers and footers of web pages, navigation elements etc. besides continuous text.

There are two basic approaches to crawling linguistically interesting data. The first, *generic* approach uses the same procedure for all documents. Its main advantage is easy implementation and wide scope in terms of text source and type. However, there are also disadvantages: data collected in this way contains more noise, has less structure and (almost) no metadata. For example, titles and subtitles are not identified, nor is the author, time and date of its publication. The second method is *target-oriented*, adjusting the implementation of crawling to individual document sources. The advantages of this approach are less noise, better

structure of collected documents and more metadata, while its weakness lies in having to adjust the crawling script for each source separately, which is time-consuming and also likely to stop working if the source modifies its platform.

The generic approach is used when building large collections of texts based on a common top-level web domain (e.g. ".si", ".uk") or the same language (e.g. slWaC corpus). The targeted approach is better suited for smaller textual collections built for specific research purposes where the structure of a text and its metadata are of key importance (e.g. the JANES corpus, described in the previous section).

Crawling typically starts with a pre-defined set of web documents, and continues with the crawler gathering new documents from hyperlinks in the existing set. The problem here is how to limit the set of collected documents to avoid gathering texts in the wrong language or genre given the purpose of the corpus compilation. There are two basic approaches when selecting which documents to crawl. The first is based on restricting URL addresses, e.g. to the domain ".si" or "med.over.net", while the second works with a list of keywords that define the target discourse domain, such as environment, tourism, cuisine etc. In this case, collecting URL addresses suitable for crawling is typically done through a search-engine API. When crawling documents for general web corpora, restricting URL addresses (e.g. for Gigafida) works best, whereas for specialised corpora keyword lists are more appropriate. Two well-known tools for the latter approach are BootCaT (Baroni and Bernardini 2004) and WebBootCaT (Baroni et al. 2006).

Web documents exist in a number of formats. The most important are *HTML* documents, which are problematic because a significant part of their content may refer to the appearance of the web page. Moreover, parts of these documents often have identical content, and in the case of textual corpora this signifies noise. Another format of documents that also contain linguistically interesting data, but is much more rarely gathered and processed, are PDF documents. The problem with collecting text from PDFs is that this format is meant for printing, so the text is encoded as characters with their positions in the page, making extraction of quality text often challenging. The following sections will thus primarily focus on describing how HTML documents are processed, while for PDF documents the extraction of content would need to be adjusted.

Another document type comes from web platforms where text is directly sent to the recipients as individual messages, similar to SMS messages or emails. By far the most well-known such platform is Twitter, a system that enables sending short messages to one's followers. Twitter also offers API scripting plugins that can be used for crawling tweets by individual authors or topics. As shown in the previous section, we gathered tweets for the JANES corpus with the TweetCat

tool (Ljubešić et al. 2014), which was purpose built for compiling tweet corpora of smaller languages. This tool, with the help of an initial language-specific word list, identifies users tweeting predominantly in the focal language (in our case Slovenian), and then via their friends and followers gradually enlarges its user base and collects their tweets together with the tweet metadata.

5.1 Procedures with generic crawling

As noted above, generic crawling is most often used when the goal is collecting a large quantity of text (more than one billion tokens) or when the human resources for collecting data are limited. The process of generic crawling for linguistically relevant data, as is also implemented in the system used for building the slWaC corpus, consists of several basic steps. The initial step is generating a *list of websites* to be crawled first. For languages with a relatively small number of speakers, such as Slovenian, this typically means a few better known websites in the language. The second step is *crawling*. Technically speaking, this step is performed by running multiple threads and searching for hyperlinks in a breadth-first approach, where the list of websites to be crawled next is updated dynamically by identifying hyperlinks from websites that have already been crawled. When a document is collected, the next step is to determine which *character encoding* is used. This piece of data should be documented in the metadata of an HTML document, yet in reality it is often missing or an incorrect encoding is declared. Determining the correct encoding system is thus mostly based on comparing the distribution of bytes in the textual part of the document to the distribution of bytes in a predetermined set of documents with known encodings.

With generic crawling, it is not possible to define the document's structure in advance, which is why a generic program, such as jusText (Pomikálek 2011) or Boilerpipe (Kohlschütter et al. 2010), has to be used. Due to its generic nature, this step creates a document structure which does not go beyond the paragraph-level nor does it collect metadata. Typically, it also does not remove all non-textual noise from the document. Next, the *language of the document* needs to be identified. This step is necessary when building a corpus, since the web is a multilingual environment. An efficient tool for this step is the langid.py script, written in Python (Lui and Baldwin 2012). The last step is *removing (near) duplicates*, since identical or nearly identical textual content is often published on multiple URL addresses. Removing (near) duplicates is most often based on calculating the intersection of word n-grams from two documents. A typical heuristic suggests that if 7-grams of two documents overlap in more than half of the cases, one can be removed as a near duplicate.

The six steps described above are mostly executed separately, which makes crawling far from optimal. The only exception is SpiderLing (Suchomel and Pomikálek 2012), which has combined the steps from crawling to language identification into an integrated process, in which individual steps communicate with each other to optimise the quantity of the crawled data and the final size of the corpus.

5.2 Procedures with target-oriented crawling

Target-oriented crawling is used when fairly little data needs to be crawled, or when there are sufficient human resources to carry out the necessary steps. This type of crawling comprises three basic steps. Specialised corpora are most often built based on a certain content and not a specific web domain. The first step is thus identifying web domains or their parts which are likely to contain plenty of sought-after content. The technical as well as legal limitations of individual sources need to be taken into account, e.g. does the website prohibit crawling (with the use of robots.txt), does it offer API scripting plugins to collect data (e.g. Twitter), and does it perhaps even allow for the entire database of texts to be downloaded (e.g. Wikipedia). The latter two options substantially ease the process of data collection, while the use of technologies such as POST and AJAX requests makes writing extractors very difficult. The next step is *crawling*, which mostly gathers all or as many documents as possible from the chosen domains. The most complicated and time-consuming is the process of writing extractors, i.e. scripts used by programmers to describe the schema of a certain type of HTML documents. This often needs to be done separately for each source, especially if its structure is very complicated, e.g. when gathering news articles and comments on these in a chronological order.

5.3 Monitor corpora

The web is particularly suitable for building monitor corpora, since its content is constantly being updated. Once the crawl platform is set up, it is simple to gather new data. This holds true for generic crawling and somewhat less so for target-oriented crawling, since individual sources may change the structure of their website, causing the original target-based extractors to stop functioning correctly.

The best tool for continuous crawling of the web are search engines, especially Google, but also local search engines (e.g. Najdi.si in Slovenia), since they are continuously trawling the web, searching for new texts. Although it is difficult

to imagine using such highly intensive processes for linguistic purposes, this can serve as the upper bound of what could be collected, and it depends on each project and the needs and abilities of its researchers how often the chosen content should be re-crawled. For researchers in lexicography, a monitor corpus would certainly be a valuable tool to detect larger and more sudden lexical changes caused by events and phenomena widely covered by media reports. and thus prompting the interest of speakers – the potential users of a dictionary. Once the first version of such dictionary is complete and made available, its authors might like to add continuous updates to its contents. In this case, building a monitor corpus and defining methods to detect new lexemes, semantic changes or changes in the characteristic context of words becomes even more important or, rather, of key importance.

6 CONCLUSION

In modern linguistics the paradigms that are used to show non-standard language versions of written communication on the Internet as a reflection of failure or pauperism of communicative abilities have somehow survived, because the analysis of language used on the Internet identifies users' ability to adapt to the electronic media or the ability to utilise the media to meet their communication needs, as they endeavour to shorten and simplify the written communication, and especially to adjust the writing to their identity (Herring 2001).

Nowadays Internet communication actively complements and changes the characteristics of the Slovenian language written for the public, to the point where a modern dictionary can no longer ignore it. In this chapter we tried to show how the web part of Gigafida can be upgraded both in its volume as well as in its topic and genre terms, and warn that such new texts should be placed into the corpus in a transparent manner (i.e. with more elaborated taxonomic categories). Some of the online genres are written in a non-standard Slovenian, which confronts corpus linguistics with an additional language technology challenge: overcoming the barriers to its automatic processing. The resources and tools with which we can help ourselves in this task are already arising in Slovenia, and different methods of crawling the Web are already being tested. The aim of the Gigafida corpus, as expanded in the proposed way, is therefore to include publicly available written production of Slovenian on the web in a broader sense; leaving the process of selection and interpretation of data from such a corpus for the needs of the dictionary to the actors in the next phase of this process, and thus to lexicographers.