Reference corpora revisited: expansion of the Gigafida corpus

Nataša Logar

Abstract

The paper discusses the expansion of the Gigafida corpus, a reference corpus of Slovenian. In order to become an even better source of language data for a new explanatory monolingual dictionary of modern Slovene, the Gigafida corpus should first be supplemented with texts from the period 2010–15 and, if possible, 1990–95. In this respect, the issues of copyright and open access to corpus texts are important, as well as issues pertaining to the criteria for the text collection process and the proportions of text types. At the end of the paper, arguments are presented for increasing the number of textbooks in the corpus, and a proposal outlined for a new taxonomy which includes topic/domain categories.

Keywords: reference corpus, Slovenian, dictionary

1 INTRODUCTION

Corpus linguistics is founded on the idea that language is primarily a social phenomenon, and as such it manifests itself exclusively in texts, which can be described and analyzed (Teubert 2005: 108). Therefore, the focus of corpus research is primarily performance (and less so or not at all competence) and observation of the language in use, which then leads to the production of theory (and not vice versa) (Kennedy 1999: 7; Leech 1992: 107). In this context, corpus linguistics differs from research approaches to language that are based on introspection, and from linguistic conclusions without evidence (Kennedy 1998: 8). Corpus linguists are not interested in which words, structures or uses of the language are possible, but rather in what is more likely to occur in a particular language, what is more frequent and typical in it, as well as what is linguistically unique or special about it. In the last three decades, corpora have become a fundamental source of data for linguistic descriptions and justifications, particularly in any modern lexicography.

"The collection of linguistic data for the dictionary must correspond to the concept, to the design of the dictionary. The relevance of the data in relation to the concept is of fundamental importance," argued Vidovič Muha at a debate part on the new dictionary of the Slovenian language, which was held at the Fran Ramovš Institute of Slovenian Language in October 2008 (Perdih 2009: 35). In the same year we were preparing specifications for the collection of corpus texts within the framework of the Communication in Slovene project (Sporazumevanje v slovenščini – SSJ),[1] with the aim of improving the previous reference corpus of Slovenian, i.e. the FidaPLUS corpus (Arhar Holdt and Gorjanc 2007), and defined the purpose of the new corpus as follows:

> Within the Communication in Slovene project there is a great number of objectives whose implementation will be based on the new corpus, including the pedagogical corpus grammar/.../ and orthography guide /.../. Slovenian lexical database will also be based on the corpus in the sense of data acquired from the corpus and its interpretations, as well as in the sense of dictionary examples. *(Korpus pisnih besedil: specifikacije /.../, December 2008: 12).*

The Gigafida corpus,[2] which was completed in 2012 (Logar Berginc et al. 2012), fully completed the pursued objectives, and with its use in preparation of the Slovenian lexical database[3] we also got the feedback on its lexical potential (Gantar 2009; 2010; 2011). Consequently, in the proposal for

1 http://eng.slovenscina.eu/
2 http://eng.slovenscina.eu/korpusi/gigafida
3 http://eng.slovenscina.eu/spletni-slovar

making a new explanatory monolingual dictionary of modern Slovene (Krek et al. 2013b), and as a starting point for the preparation of the headword list for the dictionary, it is stated that a "frequency list of the Gigafida corpus in combination with precise and relatively complex statistical analysis of the data from the corpus Kres, Gos and other databases" would be completed (ibid.: 24). The material for the new dictionary, as defined in Gliha Komac et al. (2015: 4), was very similar: "Linguistic data for making a headword list and editing of central parts of dictionary entries /.../ will come from corpus sources, mainly Gigafida, Kres, Nova beseda and partly Gos." We can therefore say once again (as in Logar et al. 2015) that the key Slovenian lexicographers in 2015 were united on the role of Gigafida and Kres in the Slovenian dictionary project, since both corpora adequately represent the lexical identity of written published Slovenian in the last 20 years (i.e. also Logar, 2014: 10 and others), although both also need to be upgraded.

The upgrading of Gigafida and Kres[4] is in the first place necessary because the last texts which were included in both were acquired on 29 May 2010, although some rather narrowly focused texts from the Internet were also obtained from the period from April 2010 to April 2011 (Logar Berginc et al. 2012: 43). Therefore, during the preparation of this paper, it should be noted that texts from books, magazines and newspapers produced less than five years ago did not exist in the Gigafida corpus. The second, perhaps more important reason for the update lies in a very modified and extended possibility of accessing the public word that changed public representation of the Slovenian language, transformed many genres that hitherto were bound only to the print, and with its associated editing processes, and brought new, specific kinds of written texts, namely the rise of new media online. And as we already wrote in Logar and Ljubešić (2013: 104):

> In defence of the necessity of building corpora – then namely corpora of *spoken* texts – Stabej and Vitez (2000) wrote: 'the fact is that the analytical picture of a certain language, which only covers the elements of written texts, is highly partial and incomplete' (79). And further on: 'if the ideal objective of a corpus-based linguistics is language comprehension, as attested in all dimensions of communication, only written corpus is insufficient' (80). The citation can be applied or it is necessarily to apply it to the texts, which a decade later are written for the 'new media'. To omit them in advance from the corpora, which represent the bases for linguistic description of a language in any dimension of communication would mean a disqualification of an important part of the language.

Krek (11. 11. 2013), during the concluding conference on the SSJ project, pointed out that during the preparation of the specifications for the Gigafida

4 Where it makes sense in continuation we refer to both.

corpus we were naturally not aware of the large increase in the use of social networks and Internet connected mobile devices that would occur after 2008, while at the same time that the reading of printed newspapers would decline. In the light of this new social reality, which has a strong influence on the language and its related descriptions, resources and technology, it is therefore necessary to rebuild reference corpora starting from good domestic and foreign practices, and plan adjustments where analysis of the corpus exposes its weaknesses.

In the following sections of the chapter we will therefore consider which segments of the Gigafida corpus should be upgraded as a priority to make it even more appropriate and relevant as a collection of linguistic data for the new explanatory monolingual dictionary of modern Slovene. Discussions on issues that require more extensive reflection (above all Internet texts) are presented in subsequent chapters of the book.

2 MODERN SLOVENE

2.1 Beginning of text collection: 1990

Language contemporaneity is a relative concept, and if we want to define the temporal dimension of texts covered by the corpus this concept necessarily requires some agreement. Consensus on the determination of the "contemporaneity" of the corpus depends on both extra- and intra-linguistic factors. Relevant for determining the starting and the finishing year of corpus texts are primarily any major changes to these. In practice, the most common reasons given for selecting the initial year of text collection (mostly rounded on a decade) are as follows:

a) time when the predecessor dictionary was published,

b) any significant socio-political changes in the language community, which brought about major changes in lexis, and

c) practical reasons, e.g. existence of electronic archives, success of the text collection process, and so on.

If we take a look at the state of modern corpora and general dictionaries of Czech and Slovak, which after 1989, due to social, political and economic events, changed or expanded their lexical funds (and even the statuses), similar to Slovenian,[5] we realise the following:

5 For example, see also a publication on Latvian by Zuicena and Miglia (2014).

a) The authors of a balanced reference corpus of Czech, prepared by the Institute of the Czech National Corpus of the Faculty of Arts in Prague, wrote in the first version of the corpus, which was made in 2000 (SYN2000,[6] followed by SYN2005 and SYN2010): "The SYN2000 is a synchronic corpus, which means that it covers contemporary Czech. Therefore it contains primarily texts that were created in 1990–1999", and the year 1990 was chosen for journalism and professional texts as a natural landmark of synchrony. The same was also true for the core part of the fiction corpus, with the exception of including books dating back further, ones that were still being reprinted and therefore affect contemporary Czech (whose author was born after 1880; for example K. Čapek and J. Hašek).[7] To this date, the most contemporary dictionary of Czech *Slovník spisovného jazyk českého* (B. Havránek et al.) is much older – and was published in four volumes in the years 1960–71, while the Institute for Czech of the Czech Academy of Sciences published it online in 2011.[8] The Institute for the Czech language is preparing a new dictionary entitled *Academic Dictionary of Contemporary Czech (Akademický slovník současné češtiny)*, but there are few publications discussing this, and these do not reveal its corpus-based methodology.[9]

b) The Ludovit Stur Institute of Linguistics of Slovak Academy of Sciences is also preparing a new dictionary, called the *Dictionary of Contemporary Slovak Language* (*Slovník súčasného slovenského jazyka*). Two volumes have already been published: the first in 2006 (A–G), the second in 2011 (H–L). It is designed as a large-scale dictionary with approximately 220,000 headwords, but its predecessor, i.e. *Dictionary of Slovak Language (Slovník slovenského jazyka)* was published four decades earlier, in the years 1959–1968 (Buzássyová 2009: 119). The primary material for the new dictionary is a lexicographical record with five million tickets and the *Slovak National Corpus*,[10] being edited since 2002 (ibid.: 124), containing texts from 1955 onwards (Šimková and Garabík 2014). In 2009 Buzássyová, who was the main editor of the dictionary (Perdih: 52), said the following:

> In theory /Slovak/ as a contemporary language is understood as from the 1940s, when Czechoslovakia split for the first time. Slovakia and its language then first took over all functions, such as the language of the arts, literature, spoken language, administration language, language for special purposes, but we do not originate from the 1940s, because that would not be realistic. /.../ We originate from the Second World War, which up to the 1960s was also covered by the previous dictionary.

6 https://ucnk.ff.cuni.cz/english/syn2000.php
7 http://wiki.korpus.cz/doku.php/cnk:syn2000
8 http://ssjc.ujc.cas.cz/
9 http://www.ujc.cas.cz/zakladni-informace/oddeleni/oddeleni-soucasne-lexikologie-a-lexikografie/akademicky-slovnik-soucasne-cestiny.html
10 http://korpus.juls.savba.sk/

The decisions made by the Czech and Slovak corpus linguists and lexicographers, and the reasons given for them, confirm a very similar argument from ten years ago on the contemporaneity of texts in the first Slovenian reference corpus, FIDA, upgraded to FidaPLUS and then to Gigafida (Gorjanc 2005: 47–48):

> The corpus FIDA tries to provide comprehensive information on modern Slovene. It tries to cover the image of today's Slovene as comprehensively as it can /.../. FIDA corpus is a synchronic corpus; it includes texts published after the year 1990 /.../. The original idea about including texts after 1980 was changed at the very beginning of the construction of the corpus changed, because of two key reasons. The first one, purely pragmatic, is related to querying the available texts in electronic form; it has been shown that the culture of electronic archives began in the second half of the nineties, so various texts should be digitized before incorporating them to the corpus. The second is related to the indexed database of the Fran Ramovš Institute of Slovenian Language that somehow provides at least basic information on the status of the language from the eighties of the last century.

And in Logar Berginc et al. (2012: 127):

> In the process of defining the time of the text collection, the collectors / of the FIDA corpus / felt that the change of the political system in Slovenia affected the language to the point that this year can be taken as a starting point for the concept of 'synchronicity' of the corpus. /.../ The corpus therefore covered the ten-year period from 1991 to 2000, with some texts from the years 1989/90.

To summarise: we put the start of text collection for the Gigafida corpus and its future upgrade for the needs of the dictionary in 1990, for the following reasons: (a) the date of publishing the last volume of the *Dictionary of Slovene Literary Language* (1970–91; DSLL); (b) socio-political changes in the late 1980s, and especially after Slovenia gained independence in 1991, that have fundamentally affected the lexical image of today's Slovene, and (c) practical reasons, i.e. the existence of the electronic archives of publishers and others.

2.2 Texts after 2010 and in the first half of the 1990s

The time period covered by the texts in the Gigafida corpus started in 1990 and finished in 2010 (print) or 2011 (Internet). The number of words per year is shown in Figure 1.

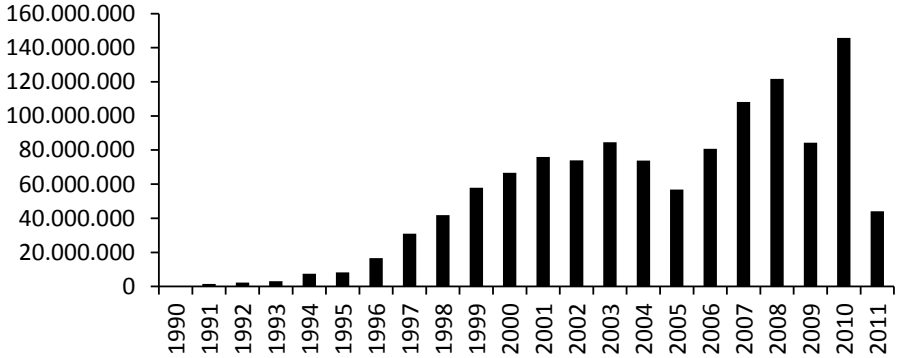

Figure 1: Number of words per year in the Gigafida corpus. Source: Logar Berginc et al. (2012: 36).

Experience shows that texts from the media are principally acquired for a year or a few years back, and less so for the current year of collection, and thus the decreases in 2005 (the year of text collection for the FidaPLUS corpus) and in 2009 are as expected. These two years can be completed, if the period before the next text collection process is not too long. An upgrade with online texts can be carried out in real-time and throughout the duration of the project, and is then stopped. In this context, crawling for the period 2012–15, in order to build up a reference corpus, remains a key step that cannot be completely replaced by another process.[11] This fact, as well as the gaps in the range of printed texts that are a consequence of excessively long periods of non-updating corpora, certainly speak in favour of longer-term infrastructure solutions, such as the Web Archive of the National and University Library[12] or the long-term financing of infrastructure projects within the Centre for the Language Resources and Technologies at the University of Ljubljana.

At the same time there are very few texts in the corpus that would enable more detailed insights into the lexical collection of Slovenian in the first half of the 1990s. For the seven-year period of 1990–96 Gigafida contains, at first glance, an extensive 22 million words, but this actually represents less than 2% of the entire corpus. If the next project of upgrading Gigafida has the budget and time needed to allow the digitisation of selected texts from this period, then this would definitely be worth considering.

11 Theoretically we could make use of the online corpus of Slovene slWaC2 (Erjavec and Ljubešić 2014), but the collection of online texts for this was not guided or controlled to the extent that is desirable with Gigafida (more on this in the next chapter).

12 http://arhiv.nuk.uni-lj.si/

3 SLOVENE IN GENERAL WRITTEN USE

3.1 Appropriateness of the corpus for general dictionary needs and purposes

We have reported several times on the text collection process for the corpora in the "FIDA series" (Gorjanc 2005: 47–53; Arhar Holdt and Gorjanc 2007; Logar Berginc and Šuster 2009; Berginc Logar et al. 2012: 21–25). Generally speaking, the key points are as follows:

a) Purpose: corpora FIDA, FidaPLUS and Gigafida were constructed in order to show a comprehensive picture of the Slovenian language, as seen in public written texts. In this sense, Gigafida as the latest corpus in series is designed to meet various linguistic research aims, but the main focus (as usually observed for the general reference corpora) is its applicability to lexical and lexicographical purposes.

b) The criteria for the collection of texts, content and documents: Gigafida as well its predecessors FIDA and FidaPLUS, used clearly drawn criteria for text collection, details of which are presented in the references, along with other, related decisions.

c) "Chasing" the general use: The criteria for text collection from the corpus FIDA onwards resulted from both reception and production. In relation with the first – if possible – this was carried out through a wider influence sieve. By doing so, we took into account objective data on readership: the National Readership Survey (newspapers, magazines); library borrowing, book awards, circulation, popularity of websites, etc. We did not take into account the collection of specialised texts (scientific) in the third stage of collection, so there are just a few of these in Gigafida. It is difficult to estimate to what extent Gigafida actually shows the general written use of the language, but the collectors never lost sight of their main goal, which was to represent this kind of use as well as possible.

A total of 77% of the words in Gigafida come from texts published in print periodicals. As we were aware that this was likely to be the case, in the SSJ project we also took samples for Kres to obtain a more balanced taxonomic share between different types of texts (Erjavec and Logar Berginc 2012).

The Gigafida corpus is therefore a large corpus and one that is heterogeneous with regard to time, genres, authors, subjects, etc. Krek and Kosem (21. 9. 2013) wrote about this as follows: "As soon as more speakers actually read certain texts (irrespective of their 'weak style'), the greater influence these texts have on their language. And so it becomes more important that lexicographers equip the content

of the dictionary database with relevant information processed from these texts for different types of dictionary users." Based on this, it appears reasonable to continue following the principle of mainly gathering texts with greater communicational influence and with a lesser (or even none existing) role for highly specialised scientific texts, when upgrading the Gigafida and Kres corpora.

3.2 The issue of a "metacorpus"

Both introductory quotations from the two proposals for the future Slovenian dictionary (Krek et al. 2013b; Gliha Komac et al. 2015) with regard to the source for the glossary and the editing of lexical entries, mention using the Gigafida corpus in combinations with Kres, Gos (a corpus of spoken Slovene), Nova beseda and other Slovenian databases. In the last decade quite an extensive selection of different corpora of Slovene has emerged (see e.g. Erjavec 2013),[13] so the question of integrating these for the purposes of dictionary editing has also naturally arisen (see also Gorjanc in Perdih 2009: 47). Or, as we wrote in Logar et al. (2015): "For the future dictionary work /.../ it is not only important the question of which corpora will be used as data collections for editing dictionary entries and why, but also the question of which corpora *will not be* used and why."

Here we speak in favour of the choice that the corpus which will be the main dataset for the general dictionary must be already made with this intention, must be carefully documented and clear in its content and structure. Only in this way will the corpus as a sample allow generalisations, which will then be published as a general-language description and regulation. With regard to the main dictionary source (in our case Gigafida together with its derivative Kres), there are of course possible combinations with other corpus resources and databases (such is, for example, the lexicographical practice in the current format of the *Great Dictionary of the Polish Language*, see Żmigrodzki 2014: 2), but we must stress that this can only happen in a way that is explained to the users of the dictionary and explicitly prescribed in the editorial process.

4 COPYRIGHT AND OPEN ACCESS

Corpora FIDA, FidaPLUS and Gigafida had legal agreements with text providers arranged in a way that it was possible to publish the corpora publicly and with

13 http://nl.ijs.si

free access. The key point here is the contractual transfer of material copyrights of the text in a way defined in Article 22 of the Slovenian Law on Copyright and Related Rights (ZASP 2007). Since the case here was accessing the texts in digital form, the holder of the rights also transmitted the rights of electronic reproduction to the providers, as set out in the first paragraph of Article 23 of the ZASP and modification rights, as set out in Article 33 ZASP:

> Article 23:
> (1) The reproduction right is the exclusive right to store the work on a material medium or another medium, directly or indirectly, temporarily or permanently, partly or in whole and in any kind of way or in any kind of form.
>
> Article 33:
> (1) The right of modification is the exclusive right that allows that a certain original work can be translated, changed for theatrical performances, musically arranged, or be modified on other ways.
>
> (2) The right from the previous paragraph also applies to cases where the original work is not unchanged but incorporated or integrated into a new work.
>
> (3) The author of the original work retains the exclusive right to use his or her work in any modified form, unless this law or contract determines otherwise.

The contract between text providers and those preparing the Gigafida corpus contained an article according to which we were allowed to use up to 10% of the text in a manner as determined by the Creative Commons licence: recognition of authorship + non-commercial + share alike, known under the denotation CC BY-NC-SA.[14] This article has enabled the composition of the corpora ccGigafida (volume of 100 million words) and ccKres (10 million words) which are accessible in the form of a database.[15]

Open access to research data from publicly funded projects was supported by all the members of OECD by signing the *Declaration on Access to Research Data from Public Funding* (OECD 2004), and Slovenia signed this in 2010 (see also *OECD Principles and Guidelines for Access to Research Data from Public Funding*).[16] The initiative with strategic documents, reports and commitments was also supported by the European Commission, the European Scientific Council, the European Federation of Academies of Sciences ALLEA and other bodies. In this respect the European Commission's recommendation on

14 https://creativecommons.org/licenses/by-nc-sa/2.5/si/legalcode
15 http://hdl.handle.net/11356/1035 in http://hdl.handle.net/11356/1034
16 http://www.oecd.org/sti/sci-tech/38500813.pdf

accessing the scientific information and their archives from 2012 is important.[17] The latter reminds EU Member States about access to publications that are the result of publicly funded research – this must be open as soon as possible, preferably immediately, and in any case not later than six months after the date of publication for the social sciences and twelve months for humanistic sciences (L194/41).[18] In the final report of the project named *Open Data – Action Plan for the Establishment of a System of Open Access to Publicly Funded Research Data in Slovenia* (2010–2013), the researchers pointed out that open research information is

> a shared responsibility of all the participants in science, which cannot be left to only one segment, for example ethical principles, but requires clearly defined obligations for individual researchers, their institutions and administrations, professional and scientific associations and other representatives of scientific community, providers of data-related services and publishers (Štebe et al. 2013: XVI).

In the future making of a reference corpus of Slovene we will have to commit to this responsibility and prepare the corpus not only for its use in a concordancer, but also in the form of "CC", which will enable domestic and foreign researchers to develop high quality, robust and useful tools for processing of natural language, in our case Slovene (Erjavec 2009: 115; Erjavec 2014). The necessity of such tools for Slovene has been pointed out on several occasions (e.g. Krek 2012b).

5 RELATED CORPORA IN TODAY'S FOREIGN LEXICOGRAPHIC PRACTICE

Table 1 shows a list of currently formatting or recently formatted general dictionaries of Finnish, Estonian, Latvian, Polish, Czech, Slovak, Dutch, German and English with the structure of the corpus, which is (was) the basis for the dictionary (if such a corpus exists).[19]

17 http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:194:0039:0043:SL:PDF

18 For more about open access see http://www.openaccess.si/

19 If for each language several general dictionaries are currently being compiled, we chose the one that is designed for web publishing; if there were several of these, as for English, the selection was random.

Table 1: List of dictionaries of nine foreign languages with the volume and contents of corpora from which they were formed or are still forming. Source: Completed and updated according to Logar (2014).

Language, dictionary, corpus	Corpus volume	Corpus contents
FINISH New dictionary of contemporary Finish / Kielitoimiston sanakirja	The dictionary is not corpus-based (Heinonen 2014).	/
ESTONIAN The Basic Estonian Dictionary (online edition in the making; Kallas et al. 2014) The Balanced Corpus of Estonian http://www.cl.ut.ee/korpused/grammatikakorpus/	15 million	• newspapers and magazines: 33% • fiction: 33% • science texts: 33%
LATVIAN Dictionary of Contemporary Latvian / Mūsdienu latviešu valodas vārdnīca www.tezaurs.lv/mlvv The Balanced Corpus of Contemporary Latvian / Līdzsvarots mūsdienu latviešu valodas tekstu korpuss www.korpuss.lv	4.5 million	• newspapers and magazines: 55% • fiction: 20% • science texts: 10% • legal texts: 8% • other: 5% • written records of parliamentary meetings: 2%
POLISH Large Dictionary of Polish Language/ Wielky słownik języka polskiego http://www.wsjp.pl/ Nacional Corpus of Polish Language / Narodowy korpus języka polskiego http://nkjp.pl/	(in the planning stage) 1.5 billion (Górski in Łazinski 2012: 33)	• newspapers, magazines and press releases: 50% • fiction: 16% • spoken texts: 10% • non-fiction: 11% • web texts: 7% • didactic texts: 2% • other: 3% • nonaligned: 1%
CZECH Akademic Dictionary of Contemporary Czech / Akademický slovník současné češtiny http://www.ujc.cas.cz/zakladni-informace/oddeleni/oddeleni-soucasne-lexikologie-a-lexikografie/akademicky-slovnik-soucasne-cestiny.html	Information about corpus-based design is not mentioned or clear.	/

Language, dictionary, corpus	Corpus volume	Corpus contents
SLOVAK Dictionary of Contemporary Slovak Language / Slovník súčasného slovenského jazyka http://slovniky.juls.savba.sk/ Slovak national corpus / Slovenský národný korpus (2013) http://korpus.juls.savba.sk/stats.html	829 million	• newspapers and magazines: 69% • non-fiction: 15% • fiction: 14% • other: 2%
DUTCH General Dutch Dictionary/ Algemeen Nederlands Woordenboek http://anw.inl.nl/search ANW Corpus / Algemeen Nederlands Woordenboek (ANW) http://anw.inl.nl/show?page=help_anwcorpus	102.5 million	• newspapers: 40% • web texts: 30% • fiction: 20% • newspapers, magazines and news portals – neologism: 5% • older texts, 1970–2000: 5%
GERMAN a) Project OWID of the Institute for German Language in Mannheim, http://www1.ids-mannheim.de/lexik/owid.html) Elexiko http://www.owid.de/wb/elexiko/start.html Elexiko-Corpus http://www.owid.de/wb/elexiko/glossar/elexiko-Korpus.html b) DWDS: A Digital Dictionary of German Language / Das Digitale Wörterbuch der Deutschen Sprache (http://www.dwds.de/) Kernkorpus[21] (http://www.dwds.de/ressourcen/kernkorpus/)	2.7 billion	• newspapers and magazines: 100% • fiction: 26% • non-fiction: 22% • scientific texts: 25% • newspapers and magazines: 27%
ENGLISH Oxford Dictionaries http://www.oed.com/ Oxford English Corpus http://www.oxforddictionaries.com/words/the-oxford-english-corpus	2.5 billion	• web texts: almost 100% (novels, non-specialised and specialised magazines, newspapers, blogs, e-mail, social networks, etc.)

20 The dictionary is based on 15 corpora, with Kernkorpus as the most important one, due to its balanced and reference structure.

The table shows that the corpora that are datasets for present and comparatively interesting dictionaries of seven foreign languages (if we overlook the Finnish and Czech) are, according to their structures, very different. If we limit ourselves to only three key categories that were most criticized in the Gigafida corpus, i.e. the small volume of fiction, large volume of journalistic texts and seemingly non-normative web text, we obtain the data in Table 2 and Picture 2 (we omit the English corpus, for which the text type composition is not publicly available, but we add data for the Czech corpus SYN2010).

Table 2: The contents of corpora of seven foreign languages and Gigafida and Kres (in %) in the categories of fiction, newspapers and magazines and web texts. Source: Completed and updated according to Logar (2014).

	Fiction	Newspapers and magazines	Web texts
Estonian	33	33	33
Latvian	20	55	0
Polish	16	50	7
Czech	40	33	0
Slovak	14	69	0
Dutch	20	45	30
German: OWID	0	100	0
German: DWDS	26	27	0
GIGAFIDA	2	77	16
KRES	17	40	20

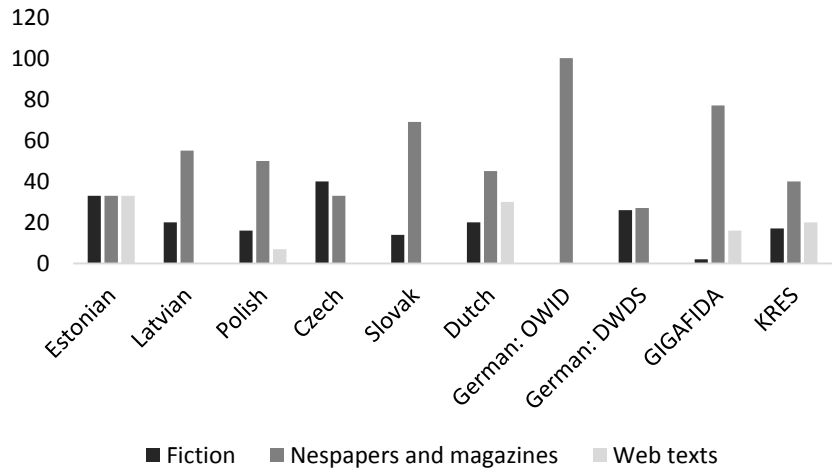

Figure 2: Contents of corpora of seven foreign languages and Gigafida and Kres (in %) in the categories of fiction, newspapers and magazines and web texts.

In Table 2 and Figure 2 we can see the following: on average more texts in the corpora come from newspapers and magazines; Gigafida has relatively little fiction, but has the largest share of journalistic texts, although the German corpus surpasses it here and the Slovak corpus is also close. Gigafida is approximately in the middle with regard to web texts. In relation to the other corpora, the components of Kres are rather average.

6 TARGETED COLLECTION OF TEXTS FOR THE PURPOSE OF THE DICTIONARY

6.1 Specialised lexis

Ledinek (2014b: 2) summarised the key issues related to the inclusion of terminology in general dictionaries as follows:

> Questions like, what is the terminology in the concrete monolingual dictionary of middle range, what will be its presumed part in the dictionary, which fields of expertise will be (in greater extent and systematically) included and what will be the way of terminology qualification (baseline) of terminology lexicon, are fundamental questions of a dictionary concept.

There is no doubt about whether to include a terminological lexicon with approximately 100,000 entries in the general dictionary or not, the question is what professional lexis and their typical text environments should be included, and in what way. The exact percentage of specialised lexis to be included in the general dictionary is debatable, but one thing is clear: to make possible any kind of collection and selection, the corpus which will form the basis for the dictionary has to be prepared in a way that it will demonstrate the state of terminological – i.e. de-terminological – lexis that is part of general language. If we leave aside the fact that such a lexicon is already reflected in the newspaper and magazine part of the corpus, as well as in the news portals part, it makes sense to follow two principles to achieve this objective when updating the Gigafida corpus:

a) the principle of *non-inclusion* of specialised texts (scientific magazines and monographs, doctoral dissertations, articles from scientific conferences, etc. precisely those that are most interesting for LSP corpora; cf. Logar, 2013: 47–52), and at the same time

b) the principle of *integration* of the popular professional works and textbooks to the level of secondary school.

We already wrote that in the final collection we avoided scientific texts, while great attention throughout the collection period after 1997 was focused on

obtaining popular professional books (manuals, guides, etc.) from various fields of human life, as well as magazines that present scientific knowledge to laymen (often younger readers). Gigafida contains almost 900 manuals from 84 different publishers and among the magazines at least 50 of them focus on some kind of expertise (e.g. motoring: *Avto Foto Market*, *Avto Magazin*, *Avtokatalog*, *Motorevija*, *Motokatalog* and *Mobil*; computing: *Connect*, *Joker*, *Moj mikro*, *Monitor*, *PC & mediji* and *Računalniške novice*). In this context it is possible to follow previous good practices and experience. The situation is different with textbooks, catalogues and didactical books, where new collections should be more systematic. Gigafida contains 103 such works, which were released by five publishers: National Education Institute Slovenia, National Examinations Centre, Rokus Klett, DZS and Ataja, but a review of included textbooks (and workbooks) shows that the scope of obligatory elementary education is covered irregularly:

- Mathematics (6 textbooks or workbooks)
- Slovene (13)
- English (1)
- History (8)
- Biology (7)
- Environmental Sciences (2)
- Physics (1)
- Chemistry (4)
- Society (4)
- Natural Sciences (1)
- Natural Sciences and Technology (1)
- Arts (1)
- Musical art (8)
- Sports (1)
- Home economics (3)

At first glance it is therefore clear that in Gigafida the obligatory school programme is not properly covered by the textbooks it includes, and there are even less works for the programs of secondary and high schools. According to the syllabus for elementary schools produced by the Ministry of Education, Science and Sport,[21] there are still missing textbooks for geography, state and civic culture, engineering and technology. From this perspective it is necessary

21 http://www.mizs.gov.si/fileadmin/mizs.gov.si/pageuploads/podrocje/os/devetletka/predmetniki/Pred_14_OS_4_12.pdf

to complete the corpus, preferably with a tendency to capture textbooks, workbooks and texts related to pupils and students of all school subjects that are part of general and vocational programs (at elementary schools, gymnasiums, and vocational secondary schools). Moreover, it would be useful to obtain information on textbooks and similar materials used for after-school extracurricular activities, particularly those with large-scale participation, and try to include this material. In this way, an upgraded Gigafida – assuming the cooperation of the text providers – would appropriately cover the terminology that almost everybody encounters during the education process. From such a corpus a collection of terms with a more comprehensive range would be extracted, and this could then be applied to the dictionary concept using a coordinated lexicographically-terminographic process.

6.2 Topic coverage

The collection of texts for reference corpora is directed by several criteria, including the diversity of text topic. In the collection of texts for the Gigafida corpus, we worked from the following list (Logar Berginc et al. 2012: 15):

- current events
- economy, politics
- education
- nature, home, pets
- people, family, men, women, children, youth
- health, food
- business, finance
- leisure, music, movie, entertainment, fashion
- sport, tourism
- culture, art
- religion, spirituality
- computing, motoring, etc.

When we used the topic modelling method to compare Gigafida with the first version of the web corpus of Slovenian slWaC (Logar Berginc and Ljubešić 2013), we found out that of twenty topics the two corpora have eight in common, seven partly in common and five different (ibid.: 92):

Characteristic to the Gigafida corpus are topics of settlements and road traffic (particularly in terms of traffic accidents), events (especially in terms of their announcement, description), television and radio programmes, individual sports and employment. In the slWaC corpus standing out are movies, music, travel and tourism, foreign policy (especially EU, Croatia), and classified ads.

From Tables 1 and 2, presented in the next chapter, we can summarise similarities and differences between the topics in the Gigafida corpus and the latest version of the slWaC$_2$ corpus, formed in 2014 (Erjavec and Ljubešić 2014):

a) Thirteen topics are common to both: human, men, woman, family life; society, other; sports; internal policy; education; finance; local politics; law; publications, culture, art; motoring; health; ICT and food.

b) Three topics are partially shared: economy (Gigafida) – economy, development (slWaC$_2$); events in the local area (Gigafida) – events (film, music, theatre) (slWaC$_2$); animals, nature, living environment (Gigafida) – living environment (slWaC$_2$).

c) Four topics are different:

- Gigafida: war, terrorism, crime; TV and radio programmes; traffic; media;
- SlWaC2: travel, tourism; online shopping; religion and internet.

Topic weaknesses of the Gigafida corpus, as indicated by this analysis, are its lack of texts about film, music and related events, travel and tourism, classified ads, external politics related to the EU, online shopping and world in general, and – surprisingly – religion. With the exception of the last one we can conclude that these are topics that in recent years have appeared quite frequently in the online media, which speaks in favour of the integration of web texts (with these topics) in the reference corpus. The analysis also confirmed the over-representation of TV and radio programmes in the Gigafida corpus (which will have to be reduced by an extended de-duplication process in the future) and the adequacy of the list of topics, which was prepared prior to the collection, although the process of updating the corpus should add the topics of law, traffic, living environment and Internet.

7 ADDITIONAL TAXONOMIC CATEGORIES

Gigafida's taxonomy is quite simple: the texts are on the first level separated into *printed* and *Internet* (see below), and then printed into *books* and *periodicals*. Literary works are divided into *fiction* and *factual texts*, and periodically printed texts

into *newspapers* and *magazines*. The category *other* is diverse (and provides only 0.67% words in the Gigafida corpus) and contains texts such as records of the meetings of the National Assembly of Republic of Slovenia, subtitles and post-production texts of the Slovenian National Television.

print
- book
  - fiction
  - factual texts
- periodical
  - newspapers
  - magazines
- other

Internet

For a general corpus search it seems that such a taxonomy is sufficient, but for lexicographical purposes it would be helpful if this would be complemented and/or further analysed. In this regard we have already indicated the need for a separate category for textbooks and similar texts, and in the next chapter we will think in this way about online texts written in non-standard Slovenian (blogs, forum posts, tweets, and comments on news portals). So far, analysis also shows that additional corpus labelling may help the lexicographers in deciding on:

- annotation of field specific lexis,
- annotation of style specific lexis.

7.1 Corpus metadata and field specific dictionary labels

Labels for field specific words or specific meanings of words, i.e. labels like *agriculture*, *motoring*, and *banking*, are closely linked with the question of terminology included in general dictionaries. If the Gigafida corpus would be at least partially labelled with topic categories, this could warn the lexicographer about a potentially sector specific meaning of the entry that he/she is editing, and at the same time such label in the corpus would allow additional sub-corpus searches. As we have already observed in Logar and Ljubešić (2013: 80), several foreign corpora have thematic categories attributed to the factual texts:

a) In the Czech National Corpus SYN2010[22] factual texts are divided into:

- religion

22 http://ucnk.ff.cuni.cz/english/syn2010.php

- law
- art
- economics
- technology
- natural sciences
- humanities and lifestyles

b) In the Croatian National Corpus[23] the layout is:

- scientific texts:
  - life sciences
  - technical science
  - biomedical sciences
  - biotechnical sciences
  - social sciences
  - humanistic science
- professional texts:
  - travel
  - reviews
  - media
  - criminology
  - sports
  - politics
  - ecology, bioethics, etc.

c) In the British National Corpus[24] under the informative texts can be found:

- world politics
- trade and finance
- art
- religion and philosophy
- leisure etc.

23 http://hnk.ffzg.hr/struktura.html
24 http://www.natcorp.ox.ac.uk/

Topic division, though not fully implemented, is, for example, also typical of the reference corpus *Oxford English Corpus*,[25] which consists of twenty parts, mostly named according to topics, e.g. computer science, environment, leisure, military, and transport. These parts are further sub-divided into sub-topics or sub-sections (sport, for example, has about 40 of these).

To achieve a complete collection of topic categories, which could be used with the texts of the upgraded Gigafida, several approaches are possible and can also be combined with one another: we could select the typology of one of the foreign corpora or rearrange the collection of topics that guided the collection of texts. A sensible approach here would be to have in sight the results of comparisons between the Gigafida and slWaC corpus obtained with the method of topic modelling and before finalising the topic scheme – to obtain key words for every corpus document with the method of TF-IDF (*Term Frequency – Inverse Document Frequency;* Salton and Buckley 1988). With the resulting topic scheme we would then manually mark the training set of documents, perform machine learning and then automatically label the corpora.

7.2 Corpus metadata and stylistic dictionary labels

The output of stylistic labels in the current version of the lexical database for Slovenian showed that the editors qualified the meanings with the following annotations in five groups (Krek et al. 2013b: 94–96):

a) **time**: *less frequent use*, *the word is very rarely used in this sense in contemporary Slovene, obsolete*[26]

b) **connotation**: *to express emphasis*, *figurative meaning*, *dissenting*, *it expresses impairment*, *pejorative*, *usually with disapproval*

c) **context**: *in journalistic jargon*, *ad texts*, *often in classified ads*, *particularly in sport*, *in Christianity*, *in a political context*

d) **pragmatics**: *as a proverb*, *with disapproval*, *euphemistically*, *usually as insult*, *rough and slightly vulgar*

d) **register**: *in very informal situations*, *in informal situations*, *in speech*, *in an informal school speech*, *informally*

To determine the *connotation* and *pragmatic labels* lexicographer must evaluate the text environment, where tools such as the Sketch Engine[27] (Kilgarriff et al.

25 http://oxforddictionaries.com/words/the-oec-composition-and-structure
26 These are just few examples from the preliminary drafting stage.
27 http://www.sketchengine.co.uk/

2004) can be a great help, while current corpus metadata may help in the time-frequency, contextual and register labels.

A. Time and frequency

Oldness or *obsolescence* of the vocabulary cannot be seen directly from corpus metadata (year of publication) since only texts issued after 1990 (mainly after 1996) are included in Gigafida. This means that the time labels can be provided by a lexicographer only on the basis of a review of the direct textual environment of the word in combination with an analysis of the frequency relationship between synonyms. On the other hand, Gigafida, with texts from a 20-year period, is relevant enough to allow reasonable annotation of labels such as *increasing use*, *decreasing use* and so on.[28] Here we must also be attentive to the frequency trend, and the fact that we should combine the increase or decrease in the frequency in a specific time period with the dispersion of sources, relative frequency depending on the number of words per year and frequency of possible synonyms. The tendency towards the transition from the labelling of timing to the labelling of frequency is in fact already seen in the preliminary set of labels in the current lexical database (e.g. *less frequent use*, *the word is rarely used*).

B. Context

Current contextual labels are diverse. They are partly linked to the analysis of a contex, which already existing corpus metadata also helps with, although to a lesser degree (e.g. lexical units from the records of the meetings of the National Assembly), and additional labelling of the corpus based on this would not help. Contextual labels are partly associated with the topic (see above, and particularly in *sport*, *Christianity*, and *political* contexts), about which we already wrote in Section 6.1.

C. Register

Register labels, the same as contextual ones, derive partly from the analysis of the context. It appears that this is primarily about identification of informal speaking situations, which can occur in all types of text, e.g. in *fiction*, in the dialogues of people in *magazines* and *newspapers*, in citations, interviews, half-literary genres or literary feuilletons. Two types of text in the Gigafida corpus were primarily spoken (records of meetings of the National Assembly and television subtitles), and both are labelled as *other* and named in the taxonomy, which directly helps a lexicographer with determining register. The third interesting source for register labels, which is also named, is the *Internet*, particularly texts

28 A chart would be most obvious in this respect.

that are to be found on news portals, and, more precisely, the texts of comments under news stories. The news sites included in the current Gigafida corpus are 24ur.com, rtvslo.si, siol.net, arhivo.com, govori.se, najdi.si (news), n-tv.si, pozareport.si, primorske.si and revija-reporter.si. The first three portals are mentioned by name, the rest have a common naming, *Internet – news*. When upgrading Gigafida with Internet texts (see next chapter) it would also be helpful to assign a separate taxonomic category to text comments as well.

8 CONCLUSION

Thirty-two researchers from eight institutions of scientific research and one publishing house cooperated in the building of the Gigafida corpus (Logar 2014: 4). The "FIDA series" corpora, which emerged over a period of almost two decades, are examples of good practice, which have followed the standards of European corpus linguistics. Therefore, when preparing the new reference corpus of Slovenian it would be good to start where we left off with Gigafida, taking into consideration the amendments which were brought into the language and text production by a new digital social reality, and the proposed improvements that were raised by the assessments of the final version of Gigafida and Kres. In this paper we did not define the structure of the future corpus of modern Slovene language. Likewise, we did not propose lists of texts that are missing in specific topics, and did not determine web sites on which it would be reasonable to perform crawling, or prepare a new taxonomy. A more concrete document must thus respond to these and related issues, such as the specification of methods used to collect texts, which is possible and sensible to prepare only when the project is approved and its time and financial frameworks are known.

The relevance of linguistic data with regard to the dictionary concept is fundamental, as we wrote in the introduction. Neither of the two existing conceptual proposals for the new dictionary of Slovenian has yet been finalised. One proposes a product "in the sense of a basic and comprehensive lexical handbook for Slovenian in the digital age", that will respectively be "conceptually, as well as from the database point of view designed completely from scratch" (Krek et al. 2013b: 20), the other will "continue the tradition of the *Dictionary of the Slovenian Language* in the sense of modern linguistic theory and in the sense of description of language use" (Gliha Komac et al. 2015: 1). Gigafida suffices to enable this baseline, but in accordance with the findings shown here and in the following chapter it can be – and should be – extended. Subsequent adjustments

will be then determined by the final dictionary concept. It will then depend on the transparency and consistency of the lexicographical process how the resulting data will be interpreted, and to what extent it will be taken into consideration, exploited or ignored.