Slovenski jezik – Slovene Linguistic Studies 16 (2024) | 149–176 | CC BY SA 4.0

**Domen Krvina** (ORCID 0000-0002-2276-1156)
ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša, Slovenija
domen.krvina@zrc-sazu.si

**Špela Petric Žižić** (ORCID 0000-0001-7451-4264)
ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša, Slovenija
spela.petric@zrc-sazu.si
DOI: https://doi.org/10.3986/16.1.07

# The relation between the composition of corpora (genre balance and representativeness) and their reliability in compiling general explanatory dictionary

This paper aims to examine the genre composition of certain Slovenian corpora as sources for lexicographic analysis (especially when compiling dictionaries such as eSSKJ, the general explanatory dictionary), particularly of the largest corpus, Gigafida 2.0 (divided into two sub-corpora: a sub-corpus of non-fiction and literary texts and a sub-corpus of journalistic texts), the Corpus of Slovenian School Texts, the Corpus of Scientific Texts of Contemporary Slovenian, as well as the KRES corpus. We argue that corpora with major discrepancy in the proportions between different text genres used as lexicographic resources do not reflect the proportions between meanings which originate in semantic extension processes. Thus, one of the largest corpora available for Slovene, Gigafida (in both versions, 1.0 and 2.0, updated in 2019), could hardly be regarded as a reference source of data for a general explanatory dictionary. This is because various journalistic texts and web texts are predominant in Gigafida, while the share of non-fiction and literary texts does not exceed 10% in total. We suggest that a corpus should be at least approximately balanced, which could in turn provide its representativeness.

KEYWORDS: Corpora, Dictionaries, Reference corpus, Representativeness, Balance, Meanings Proportion, Lexicology, Lexicography, Slovene

Namen prispevka je proučiti žanrsko sestavo nekaterih slovenskih korpusov kot virov za leksikografsko analizo (zlasti za slovarje, kot je eSSKJ, torej splošni razlagalni slovar), posebej največjega korpusa Gigafida 2.0 (razdeljenega v dva podkorpusa: podkorpus neumetnostnih in leposlovnih besedil ter podkorpus publicistike), Korpusa šolskih besedil slovenskega jezika, Korpusa znanstvenih besedil sodobne slovenščine ter korpusa KRES. V prispevku korpuse obravnavamo predvsem kot vir gradiva za izdelavo slovarjev

in sorodnih referenčnih del. Trdimo, da korpusi z večjimi odstopanji v razmerju med različnimi besedilnimi vrstami kot leksikografski viri ne odražajo razmerij med pomeni, ki so rezultat pomenotvornih procesov. Zato bi enega večjih korpusov, ki je na voljo za slovenščino, Gigafido (v obeh različicah, 1.0 in 2.0, posodobljeni leta 2019) le stežka obravnavali kot referenčni vir za splošni razlagalni slovar. V njem namreč prevladujejo različna publicistična besedila in spletna besedila, medtem ko skupni delež neumetnostnih in leposlovnih besedil ne presega 10 %. Poudarjamo, da bi korpus moral biti vsaj približno uravnotežen, kar bi posledično lahko zagotovilo tudi njegovo reprezentativnost.

KLJUČNE BESEDE: korpusi, slovarji, referenčni korpus, reprezentativnost, uravnoteženost, razmerje med pomeni, leksikologija, leksikografija, slovenščina

## 1 Introduction

The aim of this paper is to examine the genre composition of certain Slovenian corpora, particularly of the largest corpus, Gigafida (divided into a sub-corpus of non-fiction and literary texts (STVL) and a sub-corpus of journalistic texts (PUBL)) – in its current version Gigafida 2.0, and previous one, 1.0, –, the Corpus of Slovenian School Texts (*Korpus šolskih besedil slovenskega jezika*, KŠBSJ), which was made and is used especially as a source for the School Dictionary of the Slovenian Language (*Šolski slovar slovenskega jezika*), the Corpus of Scientific Texts of Contemporary Slovenian (*Korpus znanstvenih besedil sodobne slovenščine*, KZB), and the KRES corpus, as sources for lexicographic analysis. We aim to demonstrate in practice that corpora with major discrepancy in the proportions between different text genres used as lexicographic resources do not reflect the proportions between meanings within an analysed lexeme which are derived from the knowledge of lexicological theory about meaning development and semantic extension processes (cf. Atkins and Rundell 2008: 130–150, 263–309; Vidovič Muha 2013: 217; Novak 2004; Snoj 2004: 32, 77). This knowledge is (as a rule) implemented in reference lexicographic works, which are based on established lexicographic practice – as also shown in entries in authoritative explanatory dictionaries such as the Dictionary of the Slovenian Standard Language (*Slovar slovenskega knjižnega jezika*; SSKJ). In this regard, we can hardly recognize Gigafida corpus as a reference (albeit readily available) source of data for a general explanatory dictionary. This is because various journalistic texts and web texts (among them forum news comments) are distinctly predominant in Gigafida, whereas the share of non-fiction and literary texts is lower than 10% in total. In KŠBSJ

(much smaller in size and aimed at compiling a school dictionary), on the other hand, non-fiction (especially textbook) texts and texts from children's and young adult literature are predominant, which means it is a polar opposite of sorts to Gigafida in terms of its genre composition.

The markedness of journalistic texts – particularly those which exhibit also advertising characteristics,[1] which are not uncommon in the corpus Gigafida – due to the conative function taking precedence over the referential function (cf. Jakobson 1996) is a well-known fact that has been addressed in a relatively detailed manner in scholarly literature (cf. Korošec 2005). A lexical description based foremost on (recent) journalistic texts goes against general (long-term) semantic extension trends in the lexical system as reflected in descriptions of meaning in existing reference works (dictionaries such as SSKJ). Stylistic labelling relying particularly on journalistic texts (as well as genre labelling, due the lack of other genres) would be highly problematic, too, as the conative function of such texts makes them inherently stylistically marked.

We are interested in how differences in the genre composition affect the usability of a corpus in creating a relevant lexical description; which text genres are more suitable for that purpose (and what the interrelations between these genres are); how the distinct predominance of one genre, especially journalism, can affect the frequency balance of meanings within semantic relations if a lexicographer relies only on a corpus where journalism is the predominant genre, as well as how it affects the perception what is stylistically neutral and what is at least partially marked; how relying on a corpus imbalanced in terms of genre can affect dictionary descriptions; why it would make sense to pursue at least a relatively equal distribution of shares of individual genres in future corpus updates and in compiling new corpora – especially when one is not entirely sure which genre is, could or should be the most representative of the central standard register of language.[2]

---

[1] Cf. also Centa Strahovnik (2023: 24–25) on a trend in modern advertising discourse, *zaslužiš si* 'you deserve', and Gregorčič 2023 on the view of J. Habermas, who discusses the asymmetry of the operation of the media in modern society in terms of the limited engagement of users in creating or influencing content. Cf. also Vodičar 2023 about the (real) authority in the digital world, strongly influenced mostly by various marketing strategies.

[2] Cf. also Górski, Łaziński (2012: 26): Representativeness refers to a reality that exists outside the corpus. Balancing, on the other hand, is taking care to build the corpus in such a way that no component dominates the others at any level [...] The first possible rule of corpus building is not to set any criteria of representativeness, but to concentrate on building as large and diverse a corpus of randomly selected texts as possible [...].

In this context, we draw attention to the term reference corpus, which suggests that a corpus plays (or can play) the role of a reliable reference in terms of data representativeness.[3] In this regard, we highlight that a corpus not exhibiting a tendency to be at least approximately balanced, should probably not be called a reference corpus. It is worth emphasising that in this paper we analyse corpora foremost as a source of materials for compiling dictionaries (such as general explanatory dictionary eSSKJ: Dictionary of the Slovenian Standard Language 2016–) and related reference works,[4] which can indeed serve as references – under the precondition that the materials are balanced.

## 2 The methodology of building a source of materials for compiling lexicographic works

This section presents the fundamental methodological starting points for materials, which are based especially on the reflection how to provide as diverse and balanced quality material as possible so that the corpus will be as representative as possible and can thus be effectively used in dictionary compilation. A general explanatory dictionary is the core reference work in these reflections, so this overview starts with findings of Stane Suhadolnik, who led and directed the work on the Dictionary of the Slovenian Standard Language (*Slovar slovenskega knjižnega jezika*; SSKJ), an authoritative explanatory dictionary made in 1970–1991 – a recent, almost contemporary period when corpora were mostly unavailable in electronic form but were certainly judiciously structured to pursue a primary objective: to make a reference dictionary that reflects, as reliably as possible, semantic relations within lexemes and thus provides a comprehensive description of Slovenian lexical system.

### 2.1 Suhadolnik on the SSKJ corpus

On the material sources constituting the core of the corpus for compiling the dictionary in the early 1960s, Stane Suhadolnik notes in the concept of the SSKJ that this core represents approximately 2,200,000 cards that can be used in compiling a dictionary of the contemporary standard language. Among these,

[3] This term is actually used by Suhadolnik (1963: 929) as early as the 1960s to describe a dictionary, referring to what we would term the "reference status" of a reference work.

[4] "The corpus is not a replacement for any linguistic reference books" (https://korpus.sk/en/about-corpora/corpora/).

there were 700,000 reliable cards with data on ordinary words and phrases from classics and selected essayistic and popular scientific works, 400,000 good transcripts from the most recent journalism, from magazines and daily newspapers, over 100,000 cards with terminological data[5] and 900,000 cards with quotes of rarer words or phrases. All this amounted to a lexical material that made it possible to start a trial draft according to internationally valid standards (Suhadolnik 1963: 929; see also Suhadolnik 1968: 220).

To summarise, in the underlying corpus for making SSKJ, the share of non-fiction and literature was approx. 32%; the share of journalism (in the sense of anything published in recent years) was approx. 18%; the share of terminology was approx. 4.5%, and 41% of the materials were not yet categorised, consisting mostly of words occurring less frequently. The aim of the materials collected in this manner was to comprehensively demonstrate the richness of the standard language of the last half a century: the dictionary should present the vocabulary and language use of the last 60–70 years as they are reflected in the card materials and take into account the vocabulary of the classics of the previous century, modern technical terms in a secondary-school scope, as well as dialectal lexis, colloquial and jargon elements to the extent they are attested in written standard language (Suhadolnik 1968: 220). By design and size, the dictionary should serve its purpose for a number of decades and as objectively as possible present the entire central (i.e. core) vocabulary; it should not excessively fragment and map meanings and their nuances that do not actually exist in general use or represent only an emotional and sociological side of words or professional usage (Suhadolnik 1968: 223).

In a paper on language registers, Suhadolnik highlights the following with regard to dictionary descriptions, in particular:

> The central group is the most important for our study [...] This group of living, neutral words is typical for each language and the most interesting to each linguist and stylist but is sadly highly neglected because it is so widely known that nobody notices it, which is the reason even dictionaries do not address it in a satisfactory manner (Suhadolnik, Janežič 1962: 47).

[5] According to the paper, terminological lexis was selected based on pre-prepared glossaries for each field made in collaboration with experts.

It is thus apparent that Suhadolnik was aware of importance of describing the core vocabulary, first its basic meanings and only then its figurative meanings.

### 2.2 Views on building balanced corpora

The importance of genre diversity and, as far as possible, balance among the genres is also often highlighted by the compilers of various corpora, among them the compilers of the KRES corpus – this represented the basic motivation to judiciously design and build the KRES corpus as a partial adaptation of the underlying billion-word Gigafida corpus, which is distinctly imbalanced in terms of genre:

> KRES is a balanced sub-corpus sampled from Gigafida. It is key for corpora that represent the comprehensive image of a language to be large and diverse in terms of text genre. While Gigafida is such a reference corpus, it would be hard to say it is balanced as 77% of its words come from periodicals (newspapers, magazines) and only slightly over 6% from books (literature, non-fiction), for example. This composition of Gigafida is foremost the result of it including the entire FidaPLUS and everything we obtained anew under copyright contracts. This is why we have planned the 100-million KRES from the outset as Gigafida's balanced sub-corpus. (Erjavec, Logar Berginc 2012; http://www.korpus-kres.net/Support/About)

Logar Berginc et al. (2012) present the genre composition of KRES in more detail. The compilers agreed on including 17 million words from literature and 18 million words from non-fiction in KRES. To reach the agreed 18 million words, it sufficed to include 35.72% of all non-fiction texts; this share was obtained through random sampling from each title. Every newspaper out of 53 daily, weekly and free newspapers in the National Readership Survey (NRB) 2010 survey chart were also included in KRES. Magazines contribute 255,271,089 words to Gigafida, but only 20 million in KRES, which is less than 8% of the total. 20-percent share was allocated to web texts in KRES, which amounts to 20 million words, of which 8 million were allocated to texts from news portals, and 12 million to websites of institutions and enterprises (80–81).

As far as the genre composition of Gigafida is concerned, a detailed description is given for each text genre and the difficulty of obtaining such texts is highlighted. The authors note that the shares in the Gigafida taxonomy were ultimately the subjective choice of the corpus compilers. In collecting texts for

Gigafida, the data from the NRB for 2006, 2007, 2008, 2009 and 2010 was used. The most-read were some popular (free) newspapers and magazines.[6] The compilers also relied upon data from the MOSS survey (Measuring visits to websites), which was commissioned by the Slovenian Advertising Chamber.[7] In Gigafida the newspaper texts account for more than half, followed by magazines with 21%. Periodicals in total hold a 77-percent share in Gigafida. Books hold a 6-percent share in Gigafida, of which 2 percentage points of words come from literature, and 4 come from non-fiction. The authors point out that the desire to achieve 15–35% of book materials was too optimistic: approximately ten times too little literature and non-fiction was obtained to achieve 20% and 30% of the corpus, respectively. Such a deviation in the final shares can be attributed to two facts: (a) the monthly, weekly or daily production of periodicals is inherently more extensive than book production, and (b) the authors and publishers of literary works as well as everything labelled as non-fiction in the corpus are much more careful in transferring copyright than media outlets are, and approaching each individual author is more time-consuming considering the yield. Gigafida thus includes all materials the copyright was obtained and acquired for. The authors posit that if reference corpora aim to have a higher share of book materials in the future, its makers will have to be more convincing in contacting book publishers and authors; the alternative is a lower share of newspapers and magazines. Works that carry the non-fiction tag are comprised mostly of available secondary- and primary-school textbooks, manuals and guides, i.e. popular scientific and professional texts, with scientific monographs, however, appearing only sporadically. In selecting news websites, the key criterion was the number of visits. The frequency of crawling a particular website was determined intuitively; websites periodically posting listings and news about events were crawled more often, while relatively static websites were crawled less frequently (Logar Berginc et al., 2012: 21–48).

The following paragraphs present the views on how to achieve the representativeness of a corpus in the process of planning, balancing, or compiling various text corpora. Stefanowitsch notes that with large corpora composed of a broader range of web-accessible text their size is the only argument in their

[6] Such as Žurnal, Nedeljski dnevnik, Dobro jutro, Slovenske novice etc. and (magazines) Lady, Ognjišče, Motorevija, Zdravje.

[7] The 10 most visited websites included: 24.ur.com, najdi.si, siol.net, rtvslo.si, bolha.com, zurnal24, avto.net, itis.si, zadovoljna.si and enaa.com.

favour, as their creators and their users must give up pretense that they are dealing with a representative corpus. On the one hand, corpus size correlates with representativeness only to the extent that we take corpus diversity into account. On the other hand, assuming that language structure and use are not infinitely variable, size will correlate with the representativeness of a corpus at least to some extent (Stefanowitsch, 2020: 37–38). This view is shared by Corpas Pastor and Seghiri, who also point out that corpus-based studies should rely on the quality and representativeness of each corpus as foundation for producing valid results. This entails deciding on valid external and internal criteria for corpus design and compilation. A basic tenet is that corpus representativeness determines the kinds of research questions that can be addressed and the generalizability of the results obtained. Thus, the representativeness is a crucial point in the creation of a corpus, but also one of the most controversial aspects among specialists. As for the quality of the texts that are included, a system for gauging the quality of digital information through adopting an evaluation protocol should be applied to all the documents – a vast collection of texts itself is usually not sufficient as point is reached when the addition of more documents will not in practice bring anything new to the collection (Corpas Pastor, Seghiri, 2010: 111–121). This means that when compiling a corpus, one should be selective in choosing texts, so that the quality or content of the data takes equal or more precedence over issues of quantity. Corpus representativeness can be obtained by establishing coherent limits and carefully selecting textual genres for inclusion. These could be considered as external selection criteria to be established from the outset in order to ensure corpus representativeness and quality (Corpas Pastor, Seghiri, 2010: 122–135).

To summarise briefly: the readership factor should not be reduced to the mere question what is most popular.[8] In compiling a general-purpose corpus, the main principle should be the quality[9] of obtained texts and their diversity in terms of fields (and thematic diversity within fields). This should be prioritised

[8] R. L. Górski advocates that a reference corpus should reflect the readership of persons who graduated institutions of tertiary education, because these people read much more than the rest of the society. He also states that nobody knows what is the amount of text in a given newspaper which is read by an average reader (Górski 2008: 122–123).

[9] This refers both to language quality (grammatical correctness, coherent and clear syntax etc.) as well as to content quality (absence of factual errors in non-fiction texts; also avoiding advertising tendencies wherever possible).

over corpus size itself.[10] A high-quality, genre-diverse, balanced and, as far as possible, representative corpus provides a good basis to describe the core, fundamental vocabulary in its main meanings – something that, as noted by Suhadolnik, is often overlooked.

## 3 The composition of analysed text corpora

What follows is a presentation of the genre composition of, first, the Slovenian text corpora under consideration (Gigafida 2.0, KRES, KŠBSJ) and then some Slavic ones, especially Central European, which are the closest to the Slovenian language, linguistic and cultural area. We pay particular attention to the proportions between the shares of non-fiction, literature and journalism (which, in certain corpora, also encompasses most of what is defined as "web texts" in terms of their medium).

### 3.1 Gigafida, KRES, Corpus of Slovenian School Texts

| Corpus | Gigafida 2.0 | KRES | KŠBSJ | Difference Gigafida:KRES | Difference Gigafida:KŠBSJ |
|---|---|---|---|---|---|
| Size [words] | 1,1 x $10^9$ | 100 x $10^6$ | 3,9 x $10^6$ | **1** x $10^9$ | **1,096** x $10^9$ |
| Non-fiction | 3.8% | 18% | 44.1% | **–14.2 pp** | **–40.3 pp** |
| Literature | 3.5% | 17% | 50.7% | **–13.5 pp** | **–47.2 pp** |
| Journalistic texts | 64.3% | 40% | - | **24.3 pp** | - |
| Web texts | 28% | 20% | - | **8 pp** | - |
| Pupils' own texts | - | | 5.2% | | - |

Table 1: Genre distribution of texts in Gigafida 2.0 compared to KRES and KŠBSJ (shares among all texts)

The Corpus of Slovenian School Texts (*Korpus šolskih besedil slovenskega jezika*; KŠBSJ)[11] consists mainly of children's and young adult literature and of equally represented subject fields from textbook materials – which is highly relevant both to its primary use, compilation of the School Dictionary of the Slovenian Language (*Šolski slovar slovenskega jezika*; ŠSSJ),[12] and, to some extent, to a semantic analysis of texts in general. Editing work for ŠSSJ has

[10] When compiling a corpus for e.g. natural language processing (and not necessarily lexicographic analysis), its size may be more important than genre balance.

[11] On the corpus, see Ledinek et al. 2022.

[12] On the dictionary, see Petric Žižić 2020 and Petric Žižić 2022.

shown that the basic meaning (denotative, i.e. non-figurative meaning) is usually represented well both in terms of frequency and relevant collocations. Such a proportion of meanings in KŠBSJ is influenced by a large share of textbook texts, which ensures that ŠSSJ properly presents terms encountered by pupils in class. Due to the large share of literature, established figurative meanings[13] are also represented well in and proportionally to the basic meaning. Literature (especially select authors) arguably has an important impact on the semantic competence of pupils[14] – by reading such texts, they expand their own abilities to understand and form figurative meanings with regards to basic ones. Texts produced by pupils themselves have a smaller representation in the corpus,[15] but they do contribute to the presence of certain meanings typical of spoken language.

As will be shown in the SECTION 4, in Gigafida – unlike in KŠBSJ –, the basic meanings of most words, at least as far as it can be judged from the analysed sample (and can be in general supported by lexicographic experience), are often underrepresented in terms of frequency, collocations and syntactic structures. We assume this is due to Gigafida's disproportionate share of texts from particular thematic areas (e.g. sport, police blotter, automotive, food, health) of the journalistic genre, in particular. On the other hand, some meanings that are actually marginal from the point of view of general semantic relations (usually figurative meanings) but are thematically tied to journalistic areas are overrepresented – this can make it seem that the basic meaning is hardly present in use. The same can be said for some seemingly absent terminological meanings students encounter in primary school or secondary school at the latest. We can thus already note that what a corpus with a disproportionate prevalence of journalistic texts reflects is more topical, popular usage (i.e. what is written, read about; what sells) in a particular interval of time. A corpus with the predominance of non-fiction on the other hand, generally – as far as the knowledge of lexicological theory about semantic extension processes

[13] This is because children's and young adult literature are mostly not particularly experimental in linguistic terms.

[14] The evaluation of the quality of a literary work is to a certain extent intuitive, so there are no uniform criteria. However, distinguished literary works, particularly those catering to a youthful audience, are commonly acknowledged as such when the equilibrium of their tripartite functions – cognitive, ethical, and aesthetic – is maintained (Svetina 2009: 67–68).

[15] Open-access requirements resulted in legal issues, particularly relating to personal data protection (see Ledinek et al. 2022: 134).

and reference lexicographic works which follow such lexicological principles are concerned – reflects more proportionate relations between the basic and derived meanings of a given lexeme as they appear (and remain valid) over a longer interval of time.

### 3.2 Certain other Slavic (Czech, Russian, Polish, Slovak) and English (BNC) corpora

| Corpus | Russian National | Czech National (Syn2020)[16] | Polish National | Slovak National | British National[17] | Average across the 5 corpora | Gigafida 2.0 | Difference Gigafida: Average across the 5 corpora |
|---|---|---|---|---|---|---|---|---|
| Size [words] | 2,1 x $10^9$ | 100 x $10^6$ | 1,8 x $10^9$ | 1,4 x $10^9$ | 100 x $10^6$ | 1.1 x $10^9$ | 1,1 x $10^9$ | **0** |
| Non-fiction | 33.5% | 33.6% | ~30% | ~12.2% | ~50% | 37.2% | 3.8% | **–33.4 pp** |
| Literature | 40.5% | 34.7% | ~20% | 16.8% | 16.6% | 25.6% | 3.5% | **–22.1 pp** |
| Journalistic texts | 36% | 33.3% | ~50% | 71% | 18.4% | 41.6% | 64.3% (+ web texts 28%) | **22.7 pp (50.7 pp**) |

Table 2: Genre distribution of texts in national corpora of Russian, Czech, Polish, Slovak and English compared to Gigafida 2.0 (shares among all texts)

**Russian National Corpus**:[18]

Non-literary texts hold the largest share in this corpus: 59.5% (administrative 4.6%, art and culture 10.6%, history 7.2%, science and technology 9.6%, politics and society 14.2%, journalism 60.5%, everyday life 14.7%). The share of journalism in the corpus as a whole is thus 59.5% × 60.5% ~ 36%. Literary texts hold a large share: 40.5% (documentary prose 8.2%, historical prose 9.4%, fantasy 6.3%, detective stories 5%, hard-to-define prose 56.1%).

[16] This is the largest corpus in the "syn" (synchronic) group with the label "representative" (https://www.korpus.cz/kontext/corpora/corplist).

[17] In the context of corpora comparison, the shares of the parts covering the spoken language and miscellaneous are not shown here.

[18] https://ruscorpora.ru/stats

**Czech National Corpus**:[19]

Literary texts hold the largest share: 34.7% (short prose 16%, memorial prose, autobiographic prose 12 %, other prose 72%); they are closely followed by non-fiction: 33.6% (administrative 0.01%, popular scientific 39.2%, scientific 20% and scholarly literature 28%); with the share of journalism not far behind: 33.3% (leisure journalism 40%, traditional journalism 60%).

**Slovak National Corpus**:[20]

This is a corpus (prim-10.0-public-all) of a relatively small language in which the share of journalism was first around 60% and grew to roughly 70% with upgrades. Journalism thus accounts for 71%, literature for 16.8%, and non-fiction and other texts for 12.2%. A balanced sub-corpus has already been made (prim-10.0-public-vyv), with the genres above represented equally in thirds.

**Polish National Corpus**:[21]

In this corpus, journalism accounts for around 50% – but, by design, does not exceed this share, as noted by the Górski, Łaziński (2012: 29–31) – followed by non-fiction (also including administrative texts) with roughly 30%, and the share of literature is around 20%.[22]

**British National Corpus**:[23]

This corpus, the original version of which emerged around 1990 under the Oxford University Press publisher, has a prevalent, about 50% share of non-fiction (which includes scientific and popular scientific texts as well as other, e.g. administrative, essayistic, religious texts), followed by journalistic texts (newspapers and magazines) with over 18% and literature (divided into prose, poetry and drama) with 16.6%.

[19] https://www.korpus.cz/kontext/query?corpname=syn2020; https://wiki.korpus.cz/doku.php/cnk:syn2020

[20] https://korpus.sk/en/corpora-and-databases/snc-corpora/publicly-available-snc-corpora/structure-of-the-corpus-prim-10-0/

[21] https://nkjp.pl/poliqarp/

[22] The classics, which are (also) read in schools, encompass the period after 1900 – i.e. the period when the average Polish secondary school graduate does not need a dictionary to read them.

[23] https://www.english-corpora.org/bnc/

Calculating the average of fundamental genres across the 5 studied corpora reveals that non-fiction averages around 37.2%; literature amounts to 25.6% (both were also targets of the Gigafida compiler, cf. Logar Berginc et al. 2012: 33), and journalism to 41.6%. Journalism thus mostly has a moderate prevalence (or is balanced with the other genres), with the exception of the British National Corpus, but is not distinctly predominant at the expense of other genres. In fact, it would be worth pursuing such an average or proportion between the shares of fundamental genres in compiling a corpus in the first place (with consideration for the quality of the included texts). For example, the KRES corpus falls short for about 20 percentage points in non-fiction and less than 10 percentage points in literature, whereas in journalism (if not considering web texts, which are also mostly journalistic) it comes quite close.

## 4 Analysis of the individual lexemes in the compared corpora

This section presents the semantic relations within selected lexemes (mostly everyday, not specialised words) as revealed through an analysis of the corpora:

Gigafida 1.0, divided in two subcorpora – the larger PUBL[24] (consisting of texts with tags "časopisi" (newspapers), "revije" (magazines, journals), "internet" (web), "ostalo" (other)) and much smaller STVL (consisting of texts with tags "literatura" (fiction) and "stvarna besedila" (non-fiction) as well as, shared with PUBL, "ostalo" (other)) –, KRES, KŠBSJ and KZB.

The methodology is as follows:

Alongside concordance analysis we use so-called word sketches (Krek, Kilgariff 2006) with the large PUBL subcorpus of Gigafida 1.0 and comparable context search with the corpus KRES as their larger size makes context search viable. For the subcorpus STVL and the corpora KŠBSJ, KZB – due to their smaller size – we use concordance analysis only. For the lexemes with greater frequencies the calculations are based on random sample of 300 shuffled concordances. The results obtained in such a way serve as an overall lexicographic assessment of the proportions of meanings within the lexeme. The same applies to the calculated percentages of individual meanings within the lexeme in each corpus.

[24] Due to overrepresentation (90+%) of the journalistic texts using the entire Gigafida corpus for anaylsis instead of 90+% share in PUBL subcorpus would not yield much different results.

### 4.1 Proportions between meanings within a lexeme in different corpora

***Zadek*** 'abdomen, buttocks, rear': in the word sketch of the larger PUBL subcorpus of Gigafida, the figurative meaning 'rear of vehicle' is distinctly predominant, whereas the basic meanings 'rear part of body in an animal, especially an insect' and 'buttocks' are almost imperceptible (0.2%), occurring in sporadic examples (e. g. *pajkov zadek* 'spider abdomen') from popular science journals. They are much more perceptible (29,4%) in the KRES corpus (especially in coordinated structures, such *glavoprsje in zadek* 'cephalothorax and abdomen'). In KŠBSJ and STVL sub-corpus, the basic meanings are predominant, but the figurative vehicle-related meaning can also be found (9–10 %). KZB, on the other hand, displays a distinct prevalence (96%) of the basic meaning related to animals, especially insects, though there are some occurrences of the figurative meaning.

***Zadetek*** 'hit, goal, prize': in the Gigafida PUBL subcorpus as well as KRES, the meaning of a point in sports is prevalent, whereas a web search hit *iskalni zadetek* (which is now very common) or lottery prize is barely perceptible (0.1–0.6%). KŠBSJ has a relatively equal representation for the meanings of web search hit, lottery prize or sports point, whereas physical hit is not as perceptible, which also applies to KZB, where the web-search hit stands out (96%) – something not completely unexpected taking into account the type of texts included in the corpus. In the STVL sub-corpus, on the other hand, physical hit (e.g. *v oko* 'to the eye', *s topom* 'with a cannon') is more perceptible (40%), and lottery prize is less perceptible.

***Konjič*** 'small horse, wheels, horsepower, toy horse': in Gigafida PUBL subcorpus, there is a distinct prevalence (96%) of the figurative meaning 'car' or 'horsepower'. Also notable is typical of car-related texts expressive use of collocates[25], which primarily refer to an animal: *isker* 'lively', *rezgetati* 'to neigh' (*Voznik s pritiskom stopalke za plin do pločevine mobilizira 207* ***iskrih*** *konjičev, ki glasno* ***rezgetajo*** *pri 6000 vrtljajih* 'By putting the pedal to the metal, the driver mobilises 207 **lively** horses, which **neigh** loudly at 6000 rpm'), whereas the basic meaning 'animal' and the figurative meaning 'toy' are perceptible very poorly (3.5%); they are represented much better in KRES (56%). In

[25] The by-far most frequent phrase (Gigafida 1.0) *srebrnogrivi konjič* 'little silver-maned horse' refers to the proper name of an animated series and is the result of crawling frequently refreshed websites, such as TV listings, so it is irrelevant to the dictionary description. In ***lisica*** 'fox', a similar example is the phrase *zlata lisica* 'golden fox'.

KŠBSJ, KZB[26] and the STVL sub-corpus, the 'animal' meaning is predominant (82–86%), with the 'toy' meaning also perceptible. Such a difference between Gigafida and all other corpora, where the car-related meaning (and its collocates which should primarily refer to an animal) is practically imperceptible, raises the question whether to even consider this meaning (characteristic only of automotive texts) when drafting a dictionary entry in general explanatory dictionary, such as eSSKJ.

Similar examples include ***lisica*** 'fox, handcuff' (a distinct prevalence of the figurative meaning 'restraint device', especially from crime news, in Gigafida PUBL sub-corpus) and ***oven*** 'ram, Aries' (a distinct prevalence of the horoscope meaning in Gigafida PUBL sub-corpus, KRES, even the STVL sub-corpus[27]).

***Predlog*** 'proposal, suggestion, preposition': in Gigafida PUBL sub-corpus as well as KRES, there is a distinct prevalence of political and administrative context, whereas the word is perceptible rather poorly (10% in Gigafida PUBL, 19% in KRES) in everyday contexts (e.g. KŠBSJ: *Odličen predlog! Kar takoj se ga lotiva.* 'Great suggestion! Let's do it right away.') or in the 'preposition' meaning (0.3–0,7%). On the other hand, the contexts in KŠBSJ, KZB and the STVL sub-corpus are more balanced (from everyday to more formal ones, including political and administrative), and the grammatical meaning is perceptible (from 3% in STVL to 10% in KŠBSJ, and up to 25% in KZB), too.

***Koš*** 'basket, bin': in Gigafida PUBL sub-corpus, the sports meaning is distinctly prevalent (it also has distinctive collocations, e.g. *napolniti/polniti koš*, literally 'to fill the basket', i.e. 'to score points': *uspešno morajo* ***polniti koš****: tako z natančnimi meti z razdalje kot s prodori pod koš* 'they have to be successful at scoring: both by precise distance throws and breakaways'), while the meanings of an instrument to collect rubbish, laundry or carry cargo (12%) are limited mainly to individual phrases (*pleten koš* 'woven basket', *koš za smeti/odpadke* 'rubbish bin'). Those meanings are somewhat more prominent in KRES (32%) and the STVL sub-corpus (24%). The situation in KŠBSJ and KZB is similar to that in KRES and STVL, though the sports meaning only holds a minor share (3–5%).

[26] In KŠBSJ as well as KZB a significant portion of the occurrences of this lemma come from (children's) songs. Such examples show that specialised corpora might serve only as a complementary source when compiling a general explanatory dictionary.

[27] This is the result of frequently categorising non-scientific, non-professional and usually fringe fields, such as astrology, as non-fiction.

A similar example is ***vilice*** ‘fork’: in Gigafida PUBL sub-corpus, there is a distinct prevalence of the meaning referring to vehicle parts or tools (which are also advertised), especially in phrases such as *nihajne vilice* ‘swinging arm’, *(hidravlične) teleskopske vilice* ‘(hydraulic) telescopic fork’, *paletne vilice* ‘pallet fork’; as a utensil, it is used not so much as cutlery in eating (10% in Gigafida PUBL, 34% in KRES), but especially for preparing food, as influenced by numerous texts containing recipes. These technical element-related meanings are less distinct in KRES, KŠBSJ, KZB and STVL, where these meanings have a minor share (4–7%), with the basic meaning of cutlery prevalent – especially for eating and less frequently for preparing food. An example of distinct occurrence in an advertising context is ***koža*** ‘skin’ (in Gigafida PUBL sub-corpus and mostly KRES as well, there is a prevalence of collocates such as *negovati* ‘to care’, *pomirjati* ‘to soothe’, *vlažiti* ‘to moisturise’, *ščititi* ‘to protect’, *obnavljati* ‘to repair’, *gladiti* ‘to smooth’, *napeti* ‘to tighten’, *učvrstiti* ‘to firm’; *občutljiv* ‘sensitive’, *suh* ‘dry’, *masten* ‘greasy’, *razdražen* ‘irritated’, *trd* ‘hard’, *razpokan* ‘cracked’). This context of ***koža*** is substantially less (4%) present in the STVL sub-corpus and practically absent from KŠBSJ and KZB.

***Prisegati*****$_{\text{ipf}}$*****/priseči*****$_{\text{pf}}$** ‘to swear’: in Gigafida PUBL sub-corpus and mostly KRES as well, the figurative (imperfective, thus present only in ***prisegati***) meaning ‘to value, have a very good opinion of; to like to use’ (which, in SSKJ, for example, is only noted in the phraseology section as a non-standard meaning) with collocates such as *kreator* ‘creator’, *ljubitelj* ‘fan’, *zvezdnica* ‘star’, *voditeljica* ‘presenter’, *navdušenec* ‘enthusiast’, *poznavalec* ‘connoisseur’ (e.g. *Zvezdnice* ***prisegajo na*** *najrazličnejše odtenke rdeče šminke, ki odlično pristojijo njihovi polti, barvi las in oblačilom* ‘Stars swear by various shades of red lipstick, which splendidly suit their skin tone, hair colour and clothes) with the preposition *na* (e.g. *na klasiko* ‘by a classic’, *znamko* ‘trademark’, *tradicijo* ‘tradition’, *udobje* ‘comfort’, *kozmetiko* ‘cosmetics’, *eleganco* ‘elegance’, *videz* ‘appearance’, *slog* ‘style’, *lepoto* ‘beauty’) stands out so much (from 68% in KRES up to 91% in Gigafida PUBL) that it seems that ***prisegati***, despite its aspectual correlation[28] with ***priseči***, shares almost no collocates with the latter (only, for example, with the object *ljubezen* ‘love’, *zvestobo* ‘loyalty’; with the subject *predsednik* ‘president’, *domobranec* ‘Home Guard’; *pred bogom, pred*

[28] It is clear, though, that one-to-one correspondence in all collocations and especially in their frequency cannot be expected as the difference in non-categorial semantic features in an aspectual correlation is close to 0 but not actually 0. Cf. Krvina 2018.

*predsednikom* ‘before God, before the president’). It seems, for example, that a *predsednica* ‘president [f]’, *premierka* ‘prime minister [f]’, *sodnica* ‘judge [f]’, *vitez* ‘knight’ only ***priseže***$_{\textbf{pf}}$ ‘is sworn in [perfective aspect]’, never ***prisega***$_{\textbf{ipf}}$ ‘is sworn in [progressive aspect]’. In subcorpus STVL and corpora KŠBSJ, KZB and however, this figurative meaning in ***prisegati*** is proportionate (19-25%) to the basic meanings ‘to affirm that something is the truth’ and ‘to officially take up a position’.

A similar example is ***računati***$_{\textbf{ipf}}$***/izračunati***$_{\textbf{pf}}$ ‘to calculate, count’, where there is a distinct prevalence (90%) of the figurative (again, imperfective, present only in ***računati***) meaning ‘to count on’ (*resno* ‘really’, *potihem* ‘secretly’, *upravičeno* ‘legitimately’, *trdno* ‘reliably’; with the subject *selektor* ‘selector’, *trener* ‘coach’, *strateg* ‘strategist’, *prireditelj* ‘organiser’; *na podporo* ‘on support’, *pomoč* ‘help’, *uvrstitev* ‘placing’, *zmago* ‘win’, *uspeh* ‘success’, *rezultat* ‘result’, *denar* ‘money’, *uslugo* ‘favour’; *igralce* ‘players’, *kupce* ‘buyers’) in Gigafida PUBL sub-corpus and mostly (82%) KRES, too. The basic meaning is represented much more poorly (< 10–18%), so it seems that ***računati*** ‘to calculate [progressive aspect]’ and ***izračunati*** ‘to calculate [perfective aspect]’ only share the collocates (with object) *koren* ‘root’, *razdaljo* ‘distance’, *vrednost* ‘value’, *obresti* ‘interest’, *indeks* ‘index’, *porabo* ‘consumption’, *povprečje* ‘average’; (with adverb) *pravilno* ‘correctly’, *približno* ‘approximately’, *natančno* ‘precisely’, whereas the objects *razmerje* ‘ratio’, *koncentracijo* ‘concentration’, *površino* ‘area’, *hitrost* ‘velocity’, *oddaljenost* ‘distance’, *verjetnost* ‘probability’; *znesek* ‘amount’, *dohodnino* ‘income tax’ and the adjuncts *po postopku* ‘according to the procedure’, *metodi* ‘method’, *standardu* ‘standard’, *metodologiji* ‘methodology’ etc. only collocate with ***izračunati***. The situation is different in KŠBSJ, KZB and STVL, where the basic meaning, with diverse collocations (e.g. *računati ulomke* ‘to calculate fractions’; *merijo dolžine in računajo ploščine* ‘they measure lengths and calculate areas’; *koeficiente povezanosti za pojave s številčnim podatki računamo s pomočjo linearne korelacije* ‘correlation coefficients for phenomena with numerical data are calculated with a linear correlation’), is represented well, and the figurative meaning is proportionate (58–60%) in relation to it, while it is much rarer (13%) in the KŠBSJ.

***Čvrst*** ‘solid, firm’: in Gigafida and mostly KRES as well, the meaning ‘compact, elastic’, which appears in advertising texts, particularly in relation to body care (*koža* ‘skin’, *prsi* ‘breasts’, *zadnjica/ritka* ‘buttocks’, *trebuh* ‘stomach’,

*mišice* ‘muscles’, *stegna* ‘thighs’, *nohti* ‘nails’, *kosti* ‘bones’), the automotive field (*vzmetenje* ‘suspension’, *karoserija* ‘bodywork’, *podvozje* ‘undercarriage’) and in recipes (*sneg* ‘whipped egg white’, *hruška* ‘pear’, *tofu* ‘tofu’), is prominent (45–55%). The basic or derived meaning ‘full of force, strong, decisive’ appears practically only in texts on sports (*čvrsta obramba* ‘solid defence’: *V napadu igrajo zelo hitro, računam pa, da jih bomo onemogočili z našo* ***čvrsto*** *obrambo* ‘They play a very fast offence, but I’m counting on disabling them with our solid defence’). In KŠBSJ, KZB and STVL, advertising style is imperceptible (→ 0) even in phrases relating to body parts and food, and the basic meanings have a stronger presence, e.g. *bila je še čvrsta in gibčna* ‘she was still vigorous and flexible’, *stisk njegove roke je bil proti pričakovanju čvrst* ‘unexpectedly, his handshake was firm’.

To summarise briefly: search results in the STVL sub-corpus are closer to search results in KŠBSJ and KZB, whereas searching the KRES yields results somewhere between STVL/KŠBSJ/KZB and Gigafida PUBL. In KŠBSJ, KZB and STVL, the basic meaning is always perceptible well, while the representation of derived meanings depends on the inclusion of texts from a particular field in the corpus, but such meanings are mostly perceptible. If a lexicographic description for a general explanatory dictionary were to be made based purely on materials with a distinct prevalence of (mostly) advertising journalistic texts, the presentation of semantic relations within a lexeme would be more or less inadequate. Figurative meanings (e.g. to describe sports activities, automotive and other products, healthy food, body care etc.) would be distinctly prevalent.

## 5 Discussion

The above analysis shows that Slovene lacks a reference corpus which would serve as a source for comprehensive linguistic research and compiling a general explanatory dictionary, grammar and normative guide based on thorough description of the Slovenian language system.

In defining semantic relations within a lexeme, a certain theoretical basis is needed. We find the knowledge of lexicological theory about meaning development, semantic extension processes (cf. Atkins and Rundell 2008: 130–150, 263–309; Vidovič Muha 2013: 217; Novak 2004; Snoj 2004: 32, 77) to be a solid foundation for defining semantic relations within a lexeme. This knowledge

is (as a rule) implemented in reference lexicographic works which are based on established lexicographic practice. As already noted by Suhadolnik (1968), it seems that a material offering insight into processes of semantic extension (from the basic meaning into figurative meanings) is the point of reference that is desirable in lexicographic description of a lexical system.

It would therefore be worth reflecting on compiling as balanced corpus as possible which would improve the textual proportions.[29] It should be sizeable enough and while the share of periodicals (particularly weightier journalistic discussions, interviews etc., if possible) would probably be relatively large, an effort should be made that it is not overrepresented. In preserving such proportions, editorial interventions are desirable, if not necessary, to ensure texts are included according to pre-defined criteria not determined mostly by the accessibility of texts itself.

Well-defined criteria should improve current genre tagging which is often overly simplistic in terms of uniting very different text types (e.g. newspapers, journals, magazine web sites as well as some popular science publications) under a common tag with no further distinction. As far as non-fiction texts are concerned, one possibility would be to follow the example of KŠBSJ in principle: the core of non-fiction texts would consist of reviewed, mainly professional texts at the secondary school level, complemented by scientific texts (such as those collected in the KZB corpus), presumably in fundamental fields of science. Apart from this, popular science texts, instructions, various handbooks etc. should be taken into account.

Literature would be represented by quality Slovenian and foreign (semi-) literary texts. The criteria of quality – which are, as already said, always intuitive to certain extent, but generally prefer distinguished literary works which manage to maintain the equilibrium of cognitive, ethical, and aesthetic function – and readership should be both taken into account. Various lists of prize-winning literature, librarians' and similar lists of recent literature could provide some insight; world classics should probably not be omitted as well. Journalistic texts would also be selected based on the criterion of quality (daily news – reporting, discussions, articles, columns; sections marked by advertising and other texts with marketing patterns should be minimised)

[29] This includes taking into account the legal ramifications, i.e. anticipating potential legal issues and devising way to overcome them as effectively as possible (cf. Ledinek et al. 2022: 131–132).

and – regardless of their bigger production and relative ease of access – should not be overrepresented.

Such a balance between individual genres is necessary when compiling general explanatory dictionary, such as eSSKJ, as it contributes to the representativeness of the corpus data – to the extent it can be ensured when one cannot be quite sure about the actual linguistic and language-formation influence (especially in the core of language, i.e. standard language) of a genre (cf. Logar et al. 2023: 88–89). In no case should the size of the corpus take precedence over planning and compiling the corpus in a way described above. On the other hand, an excessive share of topical materials whose linguistic quality can be rather poor, provides more language innovations that are only emerging – these are represented well in the Trendi corpus (cf. Kosem et al. 2023), a valuable source of materials for the Growing Dictionary of the Slovenian Language (*Sprotni slovar slovenskega jezika*), for instance, which actually aims to capture new lexical trends (and possibly innovations).[30]

Suggestions by respondents to improve the Gigafida 2.0 corpus, obtained in a survey on the use of this corpus conducted in 2021 as part of the RSDO project, also point to the necessity of updating (and probably also compiling new) general corpora, with texts as diverse as possible in terms of genre (Logar et al. 2023: 86, 88–89). The suggestion to increase genre diversity took first place among all suggestions (Logar et al. 2023: 82).[31]

Extensive, billion-word corpora such as Gigafida (from version 1.0 onward) may be suitable for machine rather than manual analysis. However, due to their seeming primary goal to achieve the desired size, their genre composition can be seriously deficient (as noted, for example, by Górski), with a disproportionate prevalence of journalistic texts. Due to their extensive production and accessibility, such texts do enable achieving the goal of a billion-word corpus but they obscure the semantic relations within lexemes in the process (cf. Rundell, Atkins 2013: 1339). This is also reflected in the results provided by word-sketch machine analysis. In addition, due to poorer precision (and a

[30] Cf. also Krvina 2022.

[31] Judging by the overall average of ratings, ~4, the lowest-rated (~3.6) statement "the corpus offers appropriate search options" could point to this as well. The statement probably does not only refer to the range of complex (enough) search options, but very likely also to the quality of materials being searched – a user can interpret an excessive occurrence of particular search results as the inadequacy of search tools or complex search options even though the issue actually lies in the problematic genre composition of materials originating in the already mentioned distinct imbalance.

more sensational nature) of reporting, both linguistic and factual errors can occur, especially in non-peer-reviewed and thematically often relatively peripheral non-fiction texts but especially journalistic texts.[32]

If one were to pursue an effort to balance genres (it is impossible to rely only on readership data), this would involve designing the corpus (especially for compiling general explanatory dictionary) in advance so that:

1) different or at least fundamental genres (literature, non-fiction, journalism) are represented as equally as possible, despite the fact that some can be accessible in a much larger quantity than others;
2) the included texts are of the highest possible quality from the perspective of linguistic and factual correctness;
3) the texts are also represented as equally as possible in terms of the years of their creation.

This enables detecting linguistic phenomena and possible trends over a longer period; in addition, this avoids an excessively current focus – which, to a certain extent, understandably characterises dictionaries such as the *Growing Dictionary of the Slovenian Language* – by reporting on popular themes in a given year. All this entails that the design, compiling and implementation of a corpus – within technological and legal limits – must be managed by a group of editors. To obtain a sufficient number of texts, especially literary texts, it would be worth simultaneously conducting a promotion campaign to raise awareness of the importance of such texts in compiling a corpus, which would increase the willingness of both authors and publishers to provide their texts for such purposes. In any case, a high-quality composition of a corpus should be the first criterion. When considering the use of the corpus for compiling general explanatory dictionary, in particular, it's size should not be too large if that entails disrupting the proportions between fields and genres. Even though it is probably harder to achieve balance in smaller languages (as shown by Slovak, where the share of journalism in the national corpus is 60–70%, though literature still amounts to about 18%) than in large ones, it is worth pursuing exactly that – rather than mere size.

---

[32] E.g. *Nega matere in otroka*: Epiduralni anestetik **ohromi hrbtenjačo** in tako onemogoči zaznavanje bolečine 'An epidural anaesthetic **paralyses the spinal cord** and prevents feeling pain'; *Slovenske novice: 24 ur na preži*: Ta del obale skriva še eno zanimivost, dno pokrivajo goste preproge **pozejdonke**, temno zelenih **travnatih alg** 'This part of the coast reveals another interesting fact; the sea bed is covered by thick carpets of **Neptune grass**, dark green grass-like algae' (Neptune grass is not a species of algae).

The Russian and Czech national corpora feature well-balanced proportions between fields and genres. As regards language quality, a methodological commentary on the Russian National Corpus on its website highlights the importance of the adequacy of the written standard language in literary and related texts. The share of literary texts in the Czech National Corpus also amounts close to a third of the corpus; this share amounts to slightly less than a third in the Polish National Corpus (and was only a little lower in the SSKJ card index). The share of journalism, on the other hand, does not exceed a third in the Russian National Corpus, is similar in the Czech corpus (where it is further divided into "traditional", which probably means a somewhat higher level of quality, and "leisure" with a lower share) and does not exceed half of all the texts in the Polish National Corpus.

As regards the term reference corpus it seems to be sometimes used in relation to corpora, such as Gigafida, which lack balance and could hardly be considered representative – which is important for a reference corpus (cf. Atkins, Clear, Ostler 1992, Biber 1993, Gorjanc 2005). When that is the case,[33] we suggest to stop using the term reference corpus to ambiguously refer to a corpus as if it was balanced and representative.[34] The use of the phrase reference corpus is not meaningful if used every time to refer to a corpus (it seems devalued, an epithet of sorts), especially when it does not seem justified.

Górski and Łaziński note the FIDA(Plus) corpus as an outlier for its noticeably disproportionate share of journalism (and very low shares of non-fiction and literature) and the temporal limitation of texts, which are tied to the period after 1990 and especially 2000 with leaps between particular years.[35] The project

[33] "The starting point here is a clear awareness of the limited familiarity with relations between discourse and texts in a particular speech community, so analyses of these relations focus on their complexity and attempt to base criteria for the balanced nature of a corpus on this complexity" (Stefanowitsch 2020 in Logar Berginc et al. 2023: 88–89). For corpus GIGAFIDA see Corpus Compilation: Specifications, pp 3–7.

[34] Cf., for example, "Unless language structure and language use are infinitely variable (which, at a given point in time, they are clearly not), increasing the diversity of the sample will increase representativeness even if the corpus design is not strictly proportional to the incidence of text varieties or types of speakers found in the speech community. It is important to acknowledge that this does not mean that diversity and representativeness are the same thing, but given that representative corpora are practically (and perhaps theoretically) impossible to create, diversity is a workable and justifiable proxy" (Stefanowitsch 2020: 35).

[35] "One would notice that FIDA, the corpus of a small community such as Slovenes, is very poorly balanced, simply because in a country with a population of 2 million there are not enough books being written to easily create a corpus of 100 million that would not show a large imbalance between texts published in the press and in books. However, even in the case of larger communities, it may happen that certain types of texts are not very numerous, such as scientific literature in particular fields" (Górski, Łaziński 2012: 25–26).

of compiling Gigafida attempted to rectify the situation, but it seems that it pursued first and foremost a billion-word size as it seeming main objective, doing so with the accessible text production – this means mass-produced daily news and, in its tone or even altogether, advertising journalism, which has resulted in disproportions in semantic relations within a lexeme (i.e. *vilice* 'fork' first as machine part, not cutlery; *koš* 'basket' first for sports, not as a (home) accessory; *konjič* 'horse' first as a car, not an animal or toy; *nalagati* 'to load, place; to impose' first as a figurative description of obligations according to rules and regulations etc., not a concrete verb of putting something somewhere). In addition, with the domination of the conative function over the referential one (identifying objects), such mass journalism has made it difficult to judge the (un)markedness of lexis as the user gets the impression that every text has at least traces of expressive effect resulting from the conative function. Such a distinctly disproportionate prevalence of journalism with its own models of verbalisation also – not as much, but still – affects the diversity and representativeness of syntactic structures or syntactic modes of expression in general.[36]

As for the analysis of neologisms (new words and new meanings in existing words), the Trendi corpus is fairly suitable, as confirmed by the experience of its use in compiling the *Growing Dictionary of the Slovenian Language*. Due to its design, its usefulness in analysing new words surpasses that of Gigafida – even in its latest version. As regards the widest possible analyses which – if that is their purpose – encompass the whole temporal range of linguistic phenomena, there is the Metafida corpus collection, which Gigafida is part of and to the size of which it materially contributes.

In light of the experience of using the 100-million KRES, it would be worth trying whether it is possible to compile a corpus with about 150–350 million words and a balance at least close to the Polish National Corpus (the Czech and Russian corpora seem unattainable), so that the share of journalism (of the highest possible quality, consisting of daily news, reports, discussions, columns) does not exceed half of the corpus. The second half would feature a quality literature (children's, young adult, adult; Slovenian and translated) and especially adequately checked (reviewed, proofread) non-fiction primarily in the fundamental sciences,

[36] Cf. predicative modifiers in corpus Gigafida – the most common predicative modifiers are for instance *pijan* 'drunk', *mrtev* 'dead', *(ne)poškodovan* '(un)injured' (from the police blotter); *gol* 'naked', *oblečen* 'dressed' (from advertising texts); *oslabljen* 'weakened' (from sports); *vroč* 'hot', *ohlajen* 'cooled' (from recipes) (cf. Gabrovšek 2023, 118–122).

with which most people are familiarised in primary and secondary school (and later in narrower specialised fields). These two factors – an equal representation of quality texts of different genres and a size that enables the proper operation of word sketches but is also not unmanageable for manual concordance analysis – seems essential in compiling a versatile, representative and thus reference text corpus as a resource for making quality reference works (especially general explanatory dictionaries) and for independent linguistic analysis.

## References

eSSKJ: Slovar slovenskega knjižnega jezika 2016–, www.fran.si (1. 1.–31. 5. 2024).

Gigafida 2.0: Korpus pisne standardne slovenščine. https://viri.cjvt.si/gigafida/ (subcorpora PUBL and STVL available within search options: https://www.clarin.si/noske/run.cgi/first_form?corpname=gfida20_dedup;align=)

Gigafida 2.0. Corpus Compilation: Specifications. https://www.cjvt.si/gigafida/wp-content/uploads/sites/10/2019/06/Gigafida2.0_specifikacije.pdf

KRES. http://www.korpus-kres.net/ (October 2024)

Korpus šolskih besedil slovenskega jezika (KŠBSJ). Internal materials.

Korpus znanstvenih besedil (KZB). https://www.clarin.si/ske/#dashboard?corpname=kzb10

British National Corpus. https://www.english-corpora.org/bnc/

Czech National Corpus. https://www.korpus.cz/kontext/query?corpname=syn2020 (October 2024); https://wiki.korpus.cz/doku.php/cnk:syn2020 (October 2024)

Polish National Corpus. https://nkjp.pl/poliqarp/

Russian National Corpus. https://ruscorpora.ru/stats

Slovak National Corpus. https://korpus.sk/en/corpora-and-databases/snc-corpora/publicly-available-snc-corpora/structure-of-the-corpus-prim-10-0/

## Literature cited

Atkins, Sue, Clear, Jeremy, Ostler, Nicholas. 1992. Corpus Design Criteria. *Literary and Linguistic Computing* 7/1: 1–16.

Biber, Douglas. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing* 8/4: 243–257.

Centa Strahovnik, Mateja. 2023. *Čustva, človekova odnosnost in doseganje dobrega življenja*. Ljubljana: Teološka fakulteta.

Corpas Pastor, Gloria, Seghiri, Miriam. 2010. Size Matters: A Quantitative Approach to Corpus Representativeness. In: R. Rabadán, M. Fernández López, and T. Guzmán González (ed.). *Lengua, traducción, recepción en honor de Julio César Santoyo*. León: Universidad de León Área de Publicaciones: 111–145. http://hdl.handle.net/2436/622560

Gabrovšek, Dejan. 2023. Povedkov prilastek v slovenščini. *Slavistična revija* 71/2: 113–128. https://doi.org/10.57589/srl.v71i2.4108

Gorjanc, Vojko. 2005. *Uvod v korpusno jezikoslovje*. Domžale: Izolit.

Górski, Rafał L. 2008. Representativeness of a written part of a Polish general-reference corpus. Primary notes. In: B. Lewandowska-Tomaszczyk (ed.). *Corpus Linguistics, Computer Tools, and Applications – State of the Art*, Frankfurt am Main: Peter Lang. 119–123. http://nkjp.pl/settings/papers/representativeness_primary_notes.pdf

Górski, Rafał L., Łaziński, Marek. 2012. Reprezentatywność i zrównoważenie korpusu. In: A. Przepiórkowski, M. Bańko, R. L. Górski, B. Lewandowska-Tomaszczyk (ed.). *Narodowy korpus języka polskiego*. Warszawa: Wydawnictwo naukowe PWN. 25–36.

Gregorčič, Rok, 2023. Tehnološki razvoj v luči Habermasove etike diskurza. *Bogoslovni vestnik* 83/4: 911–922.

Jakobson, Roman. 1996. *Lingvistični in drugi spisi*. Ljubljana: Inštitut za humanistične študije.

Korošec, Tomo. 2005. *Jezik in stil oglaševanja*. Ljubljana: Fakulteta za družbene vede.

Kosem, Iztok, Čibej, Jaka, Dobrovoljc, Kaja, Kuzman, Taja, Ljubešić, Nikola. 2023. Spremljevalni korpus Trendi in avtomatska kategorizacija. *Slovenščina 2.0: Empirične, Aplikativne in Interdisciplinarne Raziskave* 11/1: 161–188. https://doi.org/10.4312/slo2.0.2023.1.161-188

About KRES. http://www.korpus-kres.net/Support/About (November 2023).

Krek, Kilgariff. 2006. Slovene Word Sketches. *Proceedings of 5th Slovenian/First International Languages Technology Conference*. Ljubljana. https://www.kilgarriff.co.uk/Publications/2006-KrekKilg-Ljub-SloveneWS.pdf

Krvina, Domen. 2018. *Glagolski vid v sodobni slovenščini 1. Besedotvorje in pomen*. Ljubljana: Založba ZRC. https://doi.org/10.3986/9789610500742

Krvina, Domen. 2022. The Growing Dictionary of the Slovenian Language (2014–) and Slovenian Neologisms: Study on Types of Data and Their Use. *Slovenski jezik / Slovene Linguistic Studies* 14: 117–151. https://doi.org/10.3986/sjsls.14.1.05

Ledinek, Nina, Jemec Tomazin, Mateja, Trojar, Mitja, Perdih, Andrej, Ježovnik, Janoš, Romih, Miro, Erjavec, Tomaž. 2022. Korpus šolskih besedil slovenskega jezika: zasnova in gradnja. *Jezikoslovni zapiski* 28/1: 123–137. https://doi.org/10.3986/JZ.28.1.07

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

Logar Berginc, Nataša, Gorjanc, Vojko, Arhar Holdt, Špela. 2023. Korpus Gigafida 2.0: Mnenje uporabnikov. *Jezik in Slovstvo* 68/2: 75–91.

Novak, France. 2004. *Samostalniška večpomenskost v jeziku slovenskih protestantskih piscev 16. stoletja*. Ljubljana: Založba ZRC

Petric Žižić, Špela. 2020. Tipologija razlag v Šolskem slovarju slovenskega jezika. *Slavistična revija* 68/3: 391–409. https://srl.si/ojs/srl/article/view/3875

Petric Žižić, Špela (tran.). 2022. School Dictionary of the Slovenian Language on the Franček Web Portal. *Slavistica Vilnensis,* 67/2*:* 126–140. https://orcid.org/0000-0001-7451-4264

Rundell, Michael, Atkins, Sue. 2013. Criteria for the design of corpora for monolingual lexicography. In: R. H. Gouws, U. Heid, W. Schweickard, H. E. Wiegand (eds.). *Dictionaries. An International Encyclopedia of Lexicography*. Berlin/Boston: De Gruyter Mouton. 1336–1343.

Snoj, Jerica. 2004. *Tipologija slovarske večpomenskosti slovenskih samostalnikov*. Ljubljana: Založba ZRC. https://doi.org/10.3986/9616500309

Stefanowitsch, Anatol. 2020. *Corpus linguistics: A guide to the methodology (Textbooks in Language Sciences 7)*. Berlin: Language Science Press.

Suhadolnik, Stane. 1963. Problemi slovenske leksikografije. *Sodobnost* 11/10: 926–934.

Suhadolnik, Stane, Janežič, Marija. 1962. Plasti in pogostnost leksike. *Jezik in slovstvo* 8/1–2: 45–49.

Suhadolnik, Stane. 1968. Koncept novega slovarja slovenskega knjižnega jezika. *Jezik in slovstvo* 13/7: 219–224.

Svetina, Peter. 2009. Kaj naj beremo z otroki? In: Livija Knaflič, N. Bucik (ed.). *Branje za znanje in branje za zabavo: priročnik za spodbujanje družinske pismenosti*. Ljubljana: Andragoški center Slovenije. 67–69. https://arhiv.acs.si/publikacije/Branje_za_znaje_in_branje_za_zabavo-prirocnik.pdf

Vidovič Muha, Ada. 2013. *Slovensko leksikalno pomenoslovje*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Vodičar, Janez. 2023. Avtoriteta na področju vzgoje in verovanja v digitalni dobi. *Bogoslovni vestnik* 83/4: 1035–1047.

About ZgoSLiP Project: https://www.zgoslip.si/ (November 2023).

Received November 2023, accepted December 2023.

Prispelo novembra 2023, sprejeto decembra 2023.

## Acknowledgments

The publication of article was made possible by programme Slovenski jezik v sinhronem in diahronem razvoju (P6-0038 (A)), which is financially supported by the Slovenian Research Agency (ARIS).

## Summary

### The relation between corpora composition (genre balance and representativeness) and their reliability in compiling general explanatory dictionary

The general explanatory dictionary is expected to describe the lexical system as comprehensively as possible. The knowledge of lexicological theory about meaning

development, semantic extension processes which is (as a rule) implemented in reference lexicographic works represent a sound theoretical basis for such a description. Regarding the suitable corpus for such a description to be possible, we posit that it should have adequate shares of non-fiction and literature of the highest possible quality. The share of periodicals (particularly journalistic discussions, interviews etc.) would still be relatively large; it should, however, probably not represent more than half of the corpus (as is the case, for example, in the Polish National Corpus). The corpus should be sizeable enough for (semi-)automatic analysis – but not by distorting the genre proportions through expansion. To preserve such proportions, editorial interventions are desirable, if not necessary, to ensure texts are included according to appropriate criteria, and not determined mostly by the accessibility of texts itself. The core of non-fiction would consist of reviewed, mainly professional texts at the secondary school level, complemented by scientific texts (such as those collected in the KZB corpus), preferably in fundamental fields of science. Literature would be represented by quality both Slovenian and foreign (potentially based on data collected in the ZgoSLiP project) (semi-)literary texts. Journalistic texts would also be selected based on the criterion of quality (daily news – reporting, discussions, articles, columns; sections marked by advertising and other texts with marketing patterns should be minimised) and – in spite of their bigger production and relative ease of access – should not exceed half of the corpora. Such a balance would increase the representativeness of the corpus data, even when one cannot be quite sure about the actual influence of a genre. Compiling the corpus in this way should take precedence over its size with stronger effect on the representation of long-time, representative language trends.

## Razmerje med sestavo korpusov (žanrska uravnoteženost in reprezentativnost) in njihovo zanesljivostjo pri izdelavi splošnega razlagalnega slovarja

Od splošnega razlagalnega slovarja se pričakuje čim bolj celovit opis leksikalnega sistema. Poznavanje leksikološke teorije o pomenskem razvoju in pomenotvornih procesih, ki se (praviloma) odraža v referenčnih leksikografskih delih, je dobra teoretična podlaga za tak opis. Za tak opis bi bilo smiselno zgraditi korpus z ustreznim deležem stvarne literature in leposlovja najvišje možne kakovosti. Delež periodike (zlasti novinarskih razprav, intervjujev itd.) bi bil še vedno razmeroma velik, vendar ne bi smel predstavljati več kot polovice korpusa (kot je to na primer v Poljskem narodnem korpusu). Korpus bi moral biti dovolj velik za (pol)avtomatsko analizo – vendar ne na račun izkrivljenosti žanrskih razmerij. Za ohranitev takšnih razmerij so

zaželeni, če ne celo nujni, uredniški posegi, ki zagotavljajo, da so besedila vključena v skladu z ustreznimi merili in da jih ne določa zlasti dostopnost besedil. Jedro stvarne literature bi sestavljala preverjena, predvsem strokovna besedila srednješolske ravni, dopolnjevala pa bi jih strokovna besedila (kot npr. ta, ki so zbrana v korpusu KZB), po možnosti s temeljnih področij znanosti. Leposlovje bi bilo zastopano s kakovostnimi slovenskimi in tujimi (potencialno na podlagi podatkov, zbranih v projektu ZgoSLiP) (pol)literarnimi besedili. Tudi novinarska besedila bi izbrali na podlagi merila kakovosti (dnevne novice – poročanje, razprave, članki, kolumne; besedila z oglaševalskimi prvinami v čim manjšem deležu). Kljub veliki produkciji in relativno enostavnemu dostopu publicistika ne bi smela presegati polovice korpusa. Takšno ravnovesje bi povečalo reprezentativnost korpusnih podatkov, in to kljub temu, da o dejanskem vplivu posamezne zvrsti ne moremo biti povsem gotovi. Tovrstno oblikovanje korpusa bi moralo za zastopanost reprezentativnih jezikovnih trendov imeti prednost pred velikostjo.