FidaPLUS corpus of Slovenian[1]
The New Generation of the Slovenian Reference Corpus: Its Design and Tools

Špela Arhar,[2] Vojko Gorjanc[3] and
Simon Krek[4]

1. Introduction

The paper describes the FidaPLUS corpus which is an upgrade of the Slovenian reference corpus. The corpus has been improved on various levels: size, up-to-dateness, quality of linguistic annotation (lemmatization, POS-tagging), availability and user-friendliness of the on-line concordancer. It has also been implemented in the Sketch Engine software which produces one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour. We will describe the history of the project and present the characteristics of the corpus and its tools.

2. History: the FIDA corpus

In mid 1990s, the corpus-oriented Slovenian linguistic community focused on building a 100-million corpus which was the standard size of reference corpora at the time (Erjavec, Gorjanc in Stabej 1998; Gorjanc 1999). The result was the FIDA corpus, a balanced 100-million corpus of written texts in Slovenian which was released in the year 2000.[5] The result was considered a significant achievement but there were several issues which called for immediate action to upgrade the corpus. One of them was the availability of the corpus: as the project was not funded by the state but by industrial partners in the four-partner framework (two academic, two industrial), it was not freely available to the general public but only to researchers at the two academic partners involved in the project. The second major drawback was the condition of language-processing tools for Slovenian at the time. As Slovenian is a highly inflected language, lemmatization is an extremely important process for efficient corpus research and POS-tagging presents a serious challenge. For the FIDA corpus, the NLP software was used which was available at the Amebis software company, another participant in the project[6], and which in the end produced the following result: corpus tokens included in their lexicon were supplied with the lemma and POS-tag; if one token had two or more possible lemmas (or POS-tags), all of them were attributed to the token without disambiguation. Therefore, the use of statistical tools was hindered by a number of non-lemmatized elements and non-

[1] Internet link: www.fidaplus.net.

[2] Amebis, d. o. o., Kamnik
e-mail: spela.arhar@amebis.si
[3] Department of Translation Studies, Faculty of Arts, University of Ljubljana
e-mail: vojko.gorjanc@ff.uni-lj.si
[4] Jožef Stefan Institute, Ljubljana
e-mail: simon.krek@ijs.si

[5] FIDA on the internet: www.fida.net.

[6] Amebis, d. o. o.: www.amebis.si.

disambiguated lemmas and POS-tags.[7] Nevertheless, the corpus proved to be an indispensable resource for language research purposes,[8] as well as for applicative projects such as the compilation of the first corpus-based bilingual dictionary with the Slovenian language – the Oxford-DZS Comprehensive English-Slovenian Dictionary.[9]

The upgrade of the FIDA corpus became a real option with the decision of the Slovenian Research Agency to finance the project *Language resources for Slovenian* in 2003, with Marko Stabej as the project leader, and a year later also two related projects, *The concept of corpus-based lexical and grammatical descriptions of the Slovenian language* and *The development of the Slovenian corpora network*.[10] The focal point of these projects was the upgrade of the existing FIDA corpus which was later labelled as the FidaPLUS corpus.[11] The goal was to collect texts of the type or from the publishers for which the analysis had proven to be missing in the FIDA corpus and to enlarge the corpus to the size of 300 million words. Along with the enlargement, both automatic lemmatization and the disambiguation process for multiple lemmas and POS tags should be implemented. The text collection stage of the project had been rather successful and in the end, 621 million words from various Slovenian texts were included in the corpus.

3. The FidaPLUS corpus

3.1 Corpus composition

Before the collection of texts began, a set of different criteria was established with the purpose of creating a corpus which would reflect the Slovenian discourse universe adequately. Texts were collected according to these criteria and the corpus itself is composed of subcorpora which conform with the accepted taxonomy used for text collecting.

3.1.1 Date of publication

The FidaPLUS corpus incorporates the entire collection of texts from the FIDA corpus which contained texts from 1990 till 1999. The new material dates from 1996 till 2006. Figure 1 shows the number of words in the corpus according to the date of publication. Parts of columns in black colour show the proportion of the material which originates from the FIDA corpus.

[7] See Gorjanc and Krek 2001.
[8] Gorjanc and Krek 2001, Vintar 2001, Drstvenšek 2003, Gantar 2003, Krek 2003, Kržišnik 2003, Vintar and Gorjanc 2003, Krek 2004, Gorjanc, Krek and Gantar 2005, Holz 2005, Žagar 2005, Kosem 2006.
[9] Krek et al, 2005-2006, (http://razvoj.mojdenar.si/dzsslovarji/Slovar/?id=71).
[10] For more information about the projects see: http://www.fidaplus.net/Info/Info_index_eng.html.
[11] FidaPLUS on the internet: www.fidaplus.net.

3.1.2 Linguistic proof-reading

One of the characteristics of the Slovenian sociolinguistic situation is the existence of the so-called institution of "linguistic proof-reading" (*lektoriranje* in Slovenian). It is generally considered that all texts produced by Slovenian authors with the purpose of being published should be checked for their compliance with the rules of standard Slovenian by professionals who finished the Slovene language studies and they are given relatively great freedom to modify the original. Together with the widely accepted notion that the identity of Slovenes has mostly been realized through the existence and use of language (Slovenian language community is relatively small and until 1991 it had not had its confirmation as a nation state), this situation creates an ample opportunity for people to feel their language being threatened by more widely spoken languages and for linguistic normativism. Therefore, it seemed important that the information about the secondary authorship be included in the data within corpus text heads. Figure 2 shows the percentages regarding this issue.

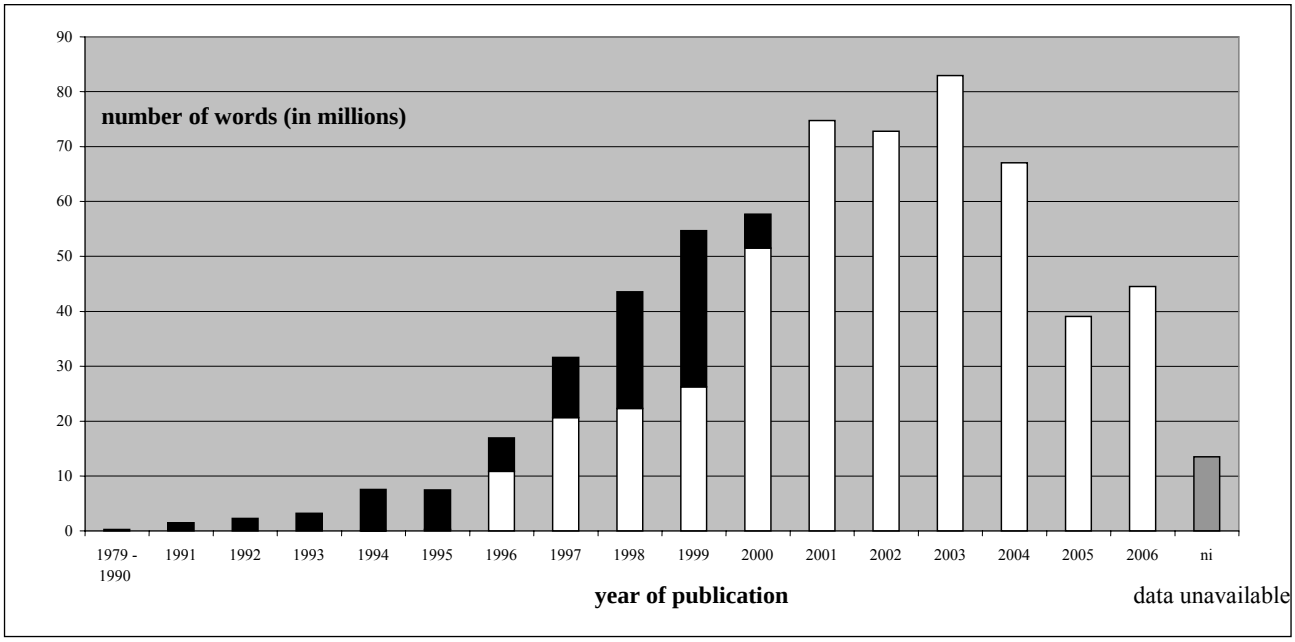

Figure 1.

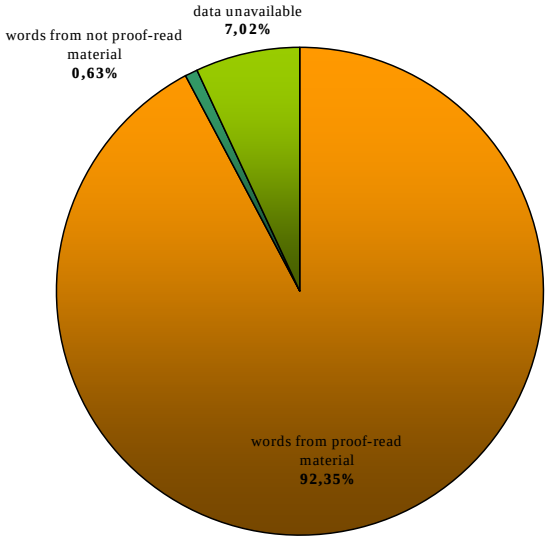

Figure 2.

3.1.3 Text variety

Texts in the FidaPLUS corpus are also categorized according to the text variety taxonomy. On the first level, they are labelled as literary and non-literary. On the second level, literary texts are characterized as prose, poetry and drama, and non-literary texts as scientific and non-scientific. Scientific texts are further divided into texts from the fields of social sciences and humanities and from natural sciences and technology. Tables 1, 2 and 3 show the exact numbers according to text variety taxonomy.

Table 1.

text type	number of words	percentage
literary	21.568.943	3,48 %
non-literary	98.871.741	96,41 %
data unavailable	709.316	0,11 %
all	**621.150.000**	100 %

Table 2.

non-literary	number of words	percentage
scientific	62.064.156	10,36 %
non-scientific	536.314.560	89,55 %
data unavailable	493.025	0,08 %
all	**598.871.741**	100 %

Table 3.

scientific	number of words	percentage
social sciences, humanities	19.331.249	31,15 %
natural sciences, technology	38.202.106	61,55 %
data unavailable	4.530.801	7,30 %
all	**62.064.156**	100 %

3.1.4 Text type

Texts are also categorized according to the text type taxonomy which includes categories such as newspapers, magazines, books, the Internet and other. The first and the second category are further subdivided according to the frequency of publication of the newspaper or the magazine. The last category (other) includes material which could not be categorized or data was unavailable, such as unpublished material or transcriptions of parliamentary talks etc. Table 4 shows the numbers.

Table 4.

type	number of words	percentage
Internet	7.682.895	1,24 %
books	54.306.387	8,74 %
newspapers	405.347.516	65,26 %
magazines	144.494.504	23,26 %
other	9.318.698	1,50 %
all	**621.150.000**	100 %

3.2 Linguistic annotation

The FidaPLUS corpus was lemmatized and POS-tagged with the software developed by the Amebis software company. The software is primarily based on the lexicon of inflectional paradigms and in the process each corpus token is compared to lexicon entries.

3.2.1 Tokens without match

If the word form was not found in the lexicon, two alternative options were considered. First, the form was compared with the list of recognized and frequent orthographic variants which are not accepted in standard Slovenian – non-standard inflection patterns,[12] unexpected agglutinations,[13] etc. The second option was

[12] Ex. *stricom* [UNCLE+inst. case] = *stricem*, the second being the standard form, lemma = *stric* [UNCLE+nom. case].
[13] Ex. *nevem* = *ne vem* [NOT KNOW+1. pers. sing.], lemma = *vedeti* [KNOW+infinitive].

unknown lemma recognition on the basis of the inflected form. This procedure is rather complicated since the different functions of a word ending have to be recognized. Thus "Americana" in "Encyclopedia Americana" can easily be confused with the genitive case, masculine gender, singular, of the non-existent lemma *american* and lemmatized as such. However, erroneous lemmatization is mainly limited to proper names of foreign-language origin.

If the process of automatic lemmatization was unsuccessful, the corpus token was left unlemmatized and the list of these elements will be used to enlarge the lexicon database in the future. The analysis of the list showed that on the top of the list there are items such as abbreviations, parts of internet addresses, non-letter combinations, parts of foreign proper names, together with some of the unrecorded neologisms (ex. *frka*, *igrovje*, *multinovela*).

3.2.2 Disambiguation

The disambiguation process was implemented if two or more lemmas were possible according to the lexicon data. An example of such cases is the double lemma of the form *padalo*. It can be either the noun *padalo* [a parachute] in nominative case, singular, or the participle of the verb *padati* [to fall], singular, neuter. The process had several stages: first, lemmas were eliminated on the basis of prefabricated rules and previously registered collocational patterns. An example of this is the collocation *pitna voda* [drinkable water]: the possible lemma *vod* [duct] was ruled out on collocational basis. Subsequently, the syntactic parser was used to disambiguate the remaining multiple lemmas. Similar process was implemented for POS-tagging.

3.2.3 New form of linguistic annotation

Both the FIDA corpus and the FidaPLUS corpus are available in the XML format and linguistic annotation in the corpus is presented in the form of attributes of the element containing one corpus token. In was decided that the information about all the possible lemmas and POS-tags would be included in the corpus, together with the disambiguated single lemma and POS tag. The new format now includes six attributes, three of them containing data about possible lemmas and the rest about POS-tag attributes (which conform with the Multext-East tagset in format).[14] In the sentence

Še vedno premočno vodi moštvo Bober I, ki je osvojilo maksimalno število točk...[15],

the verb *voditi* [to lead], is attributed the following information about possible lemmas and POS tags:

<w
lemma="voditi"
[Eng. lead(V)]
msd="Gppste--n-----n"
lemmas="voditi voda vod"
[Eng. lead(V), water(N), duct(N)]

[14] See http://nl.ijs.si/ME.
[15] *The Beaver I team which scored the maximum number of points is still in the lead...*

msds="Gppste--n-----n,Gpvsde--------n, Sozed,Sozem,Sozdi,Sozdt Sommi,Sommo"
lemmass="voditi voda vod Voda"
[Eng. lead(V), water(N), duct(N), Voda(NP)]
msdss="Gppste--n-----n,Gpvsde--------n, Sozed,Sozem,Sozdi,Sozdt Sommi,Sommo
Slzed,Slzem">
vodi</w>

Therefore, the information from the lexicon shows that the word form "vodi" has four possible lemmas (and ten corresponding POS tags): the verb "to lead", nouns "water" and "duct" and the proper name "Voda". The attribute "lemmass" always contains all possible lemmas, the attribute "lemmas" contains lemmas selected after the first level analysis and the attribute "lemma" always contains only one lemma after the second level analysis. The precision rate has not yet been quantified in exact numbers but it has been assessed to be fairly good with nouns, verbs, adjectives and numerals (around 90%) and a little less so with adverbs and closed word classes.

3.3 Concordancer

The ASP32 web concordancer was initially developed within the FIDA project and upgraded in the FidaPLUS project in terms of its functionality and design. The improvements are mainly related to the manner of presenting the information about the concordance lines, the upgrade of the statistical tool for collocation search in the corpus and a user-friendly manual for work with the concordancer.

3.3.1 New information in the concordance lines

As usual, concordance lines show minimal context of the KWIC search. In the ASP32 concordancer (Figure 1), there are two links on the left side of the concordance. One leads to the information about the text source and the other opens the full paragraph of the concordance line. The new feature of the concordancer is the link to the text source positioned in the leftmost column which itself provides the basic information about the source. With newspapers and magazines, the link shows the code of the source in the form of either the full name or a recognizable abbreviation of the source. If the text originates from sources which were not codified in this manner, their unique ID is used as the link instead. However, also in this case one can quickly recognize the broader category from which the text originates since colours are used to identify categories such as newspapers (green), magazines (blue), books (violet), the Internet (orange) and other material (grey).

3.3.2 Statistical tool

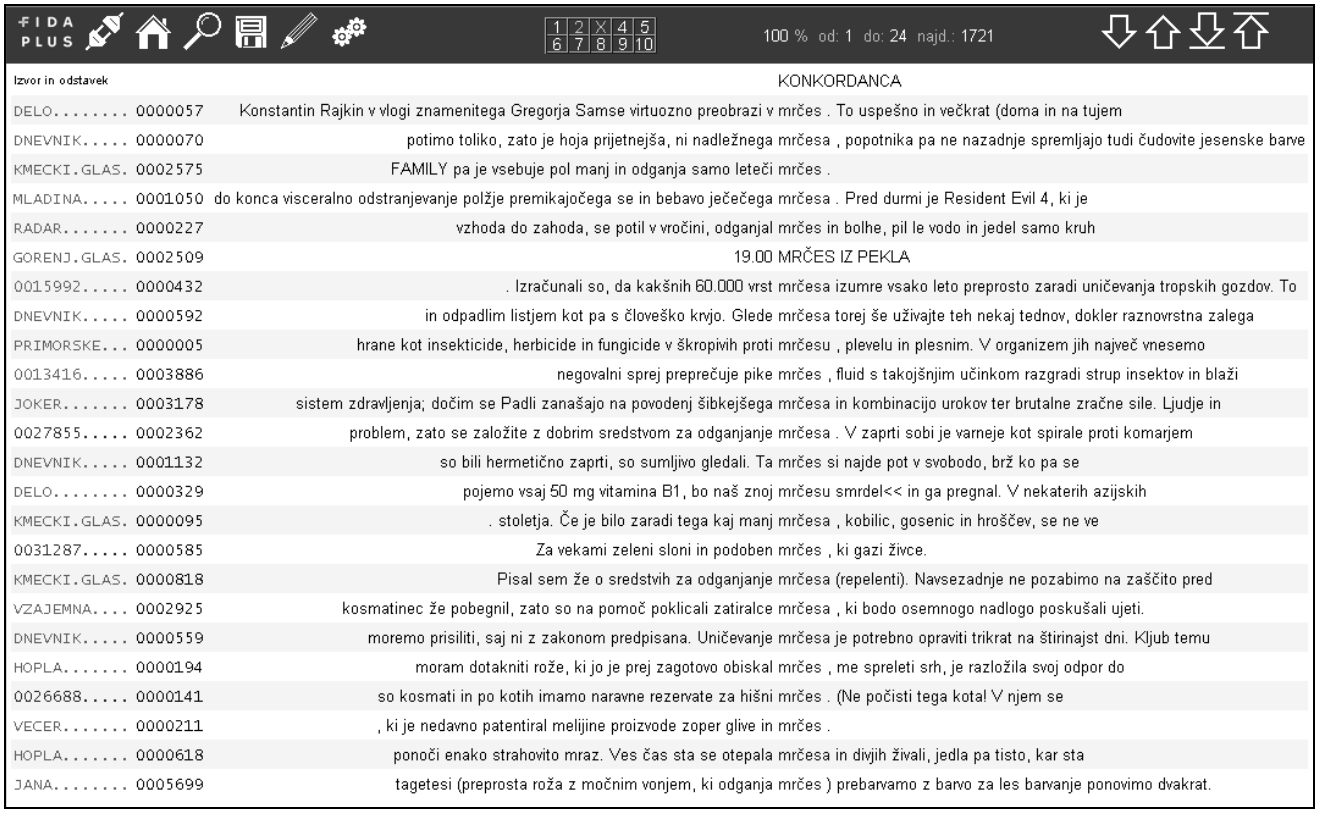

Figure 1.

The statistical tool in the previous version of the concordancer offered two statistical methods to extract collocation, the first being mutual information and the second slightly adapted MI score which put more weight on the overall number of instances of the collocates in the corpus. This was important also because of the non-lemmatized items in the FIDA corpus. In accordance with the more recent comparisons of the statistical methods indicating which statistical methods show preference for exposing low frequency items (Dunning 1993), log-likelihood was introduced as the third option in the new version of the concordancer. The renewed statistical tool enables the user to choose the frame of one to ten words either to the left or to the right side of the KWIC. Potential collocates can later be arranged according to the statistical scores, frequency or in the alphabetical order. Table 8 shows the collocates of the noun *žuželka* [insect] arranged according to the log-likelihood score.

Table 2.

ŠT.	COLLOCATE	NUMBER (of instances)	NUMBER (in the corpus)	MI SCORE	MI^3 SCORE	LL SCORE
1	pik [bite (N)]	415	5660	10.386005	27.779940	3656.750815
2	ličinka [larva (N)]	194	4071	9.764369	24.964195	1543.998837
3	čebela [bee (N)]	133	10455	7.859001	21.969566	713.019231
4	opraševati [pollinate (V)]	51	225	12.014268	23.359119	562.912130
5	hraniti [feed (V)]	152	28568	6.601439	21.097294	561.004496
6	koristen [useful (Adj)]	159	39818	6.187374	20.813140	502.138728
7	privabljati [attract (V)]	78	4888	8.185998	20.756802	452.660497
8	nadležen [pesky (Adj)]	80	5571	8.033831	20.677688	447.779343
9	loviti [hunt (V)]	114	24671	6.397985	20.063765	390.746961
10	pajek [spider (N)]	80	9279	7.297798	19.941655	368.665129
11	deževnik [earthworm (N)]	48	1186	9.528698	20.698623	366.512973
12	ptič [bird (N)]	74	8215	7.361032	19.779939	347.255385
13	pekel [hell (N)]	66	6711	7.487706	19.576494	320.890276
14	prehranjevati [feed (V)]	56	3748	8.091074	19.705784	317.784234
15	škodljiv [harmful (Adj)]	96	22939	6.255072	19.424997	311.459932
16	droben [tiny (Adj)]	109	33328	5.899361	19.435730	304.691823
17	Voden [water (Adj)]	119	41849	5.697536	19.487172	302.951024
18	pajkovec [arachnid (N)]	29	184	11.490043	21.206005	299.375793
19	nevretenčar [invertebrate (N)]	37	747	9.820113	20.239020	297.302391
20	dvoživka [amphibian (N)]	39	1359	9.032696	19.603501	271.311424

3.3.3 User manual

The lack of a user manual explaining all the complex features of the concordancer thoroughly and in a user-friendly manner was rectified with the upgrade of the concordancer. Although only available in Slovenian and thus inaccessible for the

international public, it explains the intricacies of the corpus search in three major parts: the first explaining the search methods available, the second presenting the morphosyntactic tagset used for POS tagging and the third containing a map of special characters and how they can be used in the corpus search.[16]

4. Sketch Engine

Word sketches are one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour. They were first used in the production of the Macmillan English Dictionary (Rundell 2002). At that point, they only existed for English. Now, the Sketch Engine is available, a corpus tool which takes as input a corpus of any language and corresponding grammar patterns and which generates word sketches for the words of that language. It also automatically generates a thesaurus and 'sketch differences', which specify similarities and differences between near-synonyms. The entire FidaPLUS corpus has recently been loaded into the Sketch Engine software and we shall briefly present the possibilites it offers. However, a more detailed description of the technicalities of the system is available in Kilgarriff et al. (2004), and for the Slovenian language in Krek and Kilgarriff (2006).

Word sketches are very useful for lexicographic and other purposes where one needs information about common and frequent lexical and grammatical patterns. It is based on the "gramrel" file which contains descriptions of grammatical relations for a language in a formal language which the system is able to interpret. The system uses annotation data available in the corpus, usually POS tags or syntactic categories analyzed by parsers, and presents the information in the "word sketch". Picture 1 shows the word sketch for the lemma *slep* [blind (Adj)] which is relatively simple and shows only four grammatical relations: the adjective as a modifier in a noun phrase, with a preceding verb and before a verb, the last three being mostly in the function of a nominalized adjective.

Modifiers in the first column reveal idiomatic expressions with figurative use such as *slepa pega* [blind spot], *slepa ulica* [blind alley], *slepo črevo* [vermiform appendix, also fig. a situation in which no further progress can be made], *slepi potnik, slepa potnica* [stowaway], *slepi naboj* [blank cartridge]. Some of them reveal longer constructions, such as *slepa kura* [blind hen] which uncovers the idiom *še slepa kura zrno najde* [lit. even the blind hen can find a grain – in the sense: every dog has its day] or *iti se slepe miši* [to play blind man's buff].

Coordinate structures reveal frequent set expressions such as *slepi in slabovidni* [lit. the blind and the visually impaired], *slep in gluh* [blind and deaf] which hides a longer figurative structure *slep in gluh za* [unreceptive to suggestions] besides the literal sense, and other combinations with a negative connotation.

Lexicographically relevant verbs include *voditi* which hides the structure *slepi vodi slepega* [the blind leading the blind], *ostajati* in the structure *ostajati slep za* [to remain blind for sth], *igrati* in the previously mentioned children's game blind man's buff.

[16] See Arhar 2006b.

modifies	14983	4.1
pega	490	70.13
ulica	2182	63.71
črevo	329	59.18
potnik	536	55.6
miš	229	54.56
društvo	837	41.23
mladina	294	40.46
kura	91	40.03
naboj	117	37.57
tir	127	37.35
tipkanje	35	33.66
center	537	33.21
potnica	27	32.31
pokorščina	26	31.85
vera	117	31.5
zaljubljenost	26	28.53
poslušnost	23	28.38
intelektualec	48	28.1
otrok	346	27.08
Sanja	26	26.82
rokav	47	25.8
Loka	97	24.55
sovraštvo	41	24.49
človek	328	24.38
vodnik	66	24.36

coord	5040	2.6
slaboviden	3541	125.24
gluh	396	68.17
hrom	15	31.74
gluhonem	11	26.34
nem	17	22.14
neumen	12	18.2
debel	21	17.45
invaliden	10	16.77
gol	15	12.77
star	34	8.86
da	19	8.44
pa	15	7.7
ne	10	5.26
se	18	4.74
on	11	2.2
brez	11	2.04

prec_verb:	643	0.4
voditi	37	22.35
ostajati	25	20.33
pomagati	28	19.72
postati	25	19.31
igrati	27	19.22
roditi	11	17.55
imenovati	15	15.43
omogočati	17	15.39
potrebovati	16	13.34
delati	20	12.74
uporabljati	14	12.13
narediti	10	10.87
predstavljati	10	10.13
iti	11	5.66

post_verb:	329	0.2
pričati	14	22.34
meriti	13	18.93
omogočati	18	18.77
zanimati	11	15.48
potrebovati	12	13.49
videti	11	12.02

5. Conclusion

Dynamic growth of the reference corpus remains one of the important tasks of the Slovenian corpus linguistics community. This has become evident after the FIDA corpus was released – soon it had shown signs of aging which was remedied after the FidaPLUS project had been finished but only stable financing would ensure its continuous relevance. Together with the continuous collecting of new text material, NLP tools for more accurate linguistic annotation and tools for automatized corpus analysis should be developed.

The project of building a spoken corpus of Slovenian is still in its very beginnings. A pilot project has been finished (Zemljarič Miklavčič: 2006) and represents a major step on the way but spoken corpus should become one of the priorities in the corpus-oriented community. And finally, after the first important applicative (dictionary) projects, one would hope for a more general corpus-based description of the Slovenian language. With the reference corpus at hand, it would be inadmissible to use out-dated methodologies in its design and development.

Bibliography

Arhar, Š. (2006b) Kaj početi z referenčnim korpusom FidaPLUS. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta. Elektronski vir. URL: http://www.fidaplus.net (accessed: 18 May 2007).

Drstvenšek, N. (2003): 'Vloga besedilnega korpusa pri postavitvi geselskega članka v enojezičnem slovarju'. Jezik in slovstvo 48/5, pp. 65–81.

Dunning, T. (1993) 'Accurate Methods for the Statistics of Surprise and Coincidence'. Computational Linguistics, 19/1, pp. 61–74.

Erjavec, T., Gorjanc, V. in Stabej, M. (1998) Korpus FIDA, in Jezikovne tehnologije za slovenski jezik /Language Technologies for the Slovene Language, pp. 124–127. Ljubljana: Institut Jožef Stefan.

Gantar, P. (2003) Stalnost in spremenljivost frazema v slovarju, in Stanisław Gajda in Ada Vidovič Muha (eds.) Współczesna polska i słoweńska sytuacja językowa, pp. 209−223. Opole: Uniwersytet Opolski, Instytut Filologii Polskiej/Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.

Gorjanc, V. (1999) Korpusi v jezikoslovju in korpus slovenskega jezika FIDA, in 35. seminar slovenskega jezika, literature in kulture, pp. 47–59.

Gorjanc, V. in Krek, S. (2001): A corpus-based dictionary database as the source for compiling Slovene-X dictionaries, in Proceedings of the COMPLEX 2001 6th Conference on Computational Lexicography and Corpus Research, pp. 41−47.

Gorjanc, V., Krek, S. in Gantar, P. (2005): 'Slovenska leksikalna podatkovna zbirka'. Jezik in slovstvo 50/2, pp. 3–19.

Holz, N. (2005): 'Mesto Velikega slovarja tujk v slovenski leksikografiji'. Jezik in slovstvo 50/1, pp. 87–99.

Kosem, I. (2006): 'Definicijski jezik v Slovarju slovenskega knjižnega jezika s stališča sodobnih leksikografskih načel'. Jezik in slovstvo 51/5, pp. 25–45.

Krek, S. (2003) 'Sodobna dvojezična leksikografija'. Jezik in slovstvo 48/1, pp. 45–60.

Krek, S. (2004) 'Slovarji serije COBUILD in formalizacija definicijskega jezika'. Jezik in slovstvo 49/2, pp. 3–16.

Krek, S. et al. (2006) The Oxford-DZS comprehensive English-Slovenian dictionary. Ljubljana: DZS.

Kržišnik, E. (2003): Novosti v slovenski frazeologiji, , in Stanisław Gajda in Ada Vidovič Muha (eds.) Współczesna polska i słoweńska sytuacja językowa, pp. 191−208. Opole: Uniwersytet Opolski, Instytut Filologii Polskiej/Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.

Rundell, M. (ed.) (2001). Macmillan English Dictionary for Advanced Learners. Macmillan Education.

Vintar, Š. (2001): 'Using parallel corpora for translation-oriented term extraction'. Babel 47/2, pp. 121−32.

Vintar, Š. in Gorjanc, V. (2003): 'Identifying markers of semantic relations in Slovene'. Strani jezici 1-2, pp. 37−44.

Zemljarič Miklavčič, J. (2006) Korpus govorjene slovenščine, in Tomaž Erjavec in Jerneja Gros (eds.) Jezikovne tehnologije/Language Technologies, pp. 124–127. Ljubljana: Institut Jožef Stefan.

Žagar, M. (2005): 'Determinologizacija (na primeru terminologije fizike)'. Jezik in slovstvo 50/2, pp. 35–48.