Corpus-Linguistic Topics

UDC 811.163.6'373
Vojko Gorjanc
Faculty of Arts, Ljubljana

CORPUS LINGUISTICS AND LEXICAL DESCRIPTIONS OF THE SLOVENIAN LANGUAGE

The paper presents a brief overview of the history of the corpus approach in Slovenian language studies and the existing corpora of the Slovenian language. These corpora have provided an incentive for a series of thorough linguistic studies, both monolingual and contrastive; at the same time they are becoming an indispensable part of general linguistic research, especially in the field of lexical or lexicosemantic studies. In the second part of the paper, a case study illustrates one of the procedures in lexical corpus analysis: using selected examples, we demonstrate how it is possible to track changes in the lexis of the Slovenian language in the last decade of the twentieth century.

V članku na kratko predstavimo zgodovinsko ozadje korpusnega pristopa v slovenističnem jezikoslovju, ob tem pa tudi obstoječe korpuse slovenskega jezika. Ti so v bili za jezikoslovje v slovenskem prostoru pobudni za vrsto celovitih korpusnih študij, tako enojezičnih kot tudi kontrastivnih, hkrati pa postajajo vse bolj nepogrešljiv del jezikoslovnega raziskovalnega dela sploh, predvsem ko gre za leksikalne oz. leksikalnopomenske študije. V drugem delu s študijo primera prikažemo enega od postopkov leksikalne korpusne analize: z izbranimi zgledi pokažemo na možnosti sledenja spremembam leksike slovenskega jezika v zadnjem desetletju prejšnjega stoletja.

Key words: corpus linguistics, lexical semantics, Slovenian corpora
Ključne besede: korpusno jezikoslovje, leksikalno pomenoslovje, korpusi slovenščine

1 Introduction

In the last decade, corpus linguistics has established itself as a separate research starting point, strictly empirical in nature, in which language is explored exclusively on the basis of texts which form a universe of discourse and are collected in corpora for research purposes. Corpus linguistics focuses primarily on the meaning which manifests itself as language use (Teubert 1999). Within this framework, the starting point for contemporary lexical descriptions is the analysis of large samples of materials collected with a purpose and the empirical analysis of actual samples of language use (Biber et. al. 1998: 5, 9–10). These characteristics cannot be found in older pre-computer corpora (Čermák 2002: 265). Setting standards, based on the analysis of discourse space, for including texts in corpora contributes in an important way to the quality of the language data found in a corpus. In this way, it is possible to establish a distinction between the typical and the special/individual, i.e. the recognition of the central and the peripheral language phenomena, and the observation of their distribution in different texts (Gorjanc, Krek and Gantar: 2001: 4), among other things by comparing their times of creation. In Slovenia, different types of corpora have emerged in the past few years thus establishing the field of corpus linguistics as a

separate research starting point. Corpora were, of course, a necessary prerequisite for such a development, but in the last few years a number of corpus-based linguistic studies have been carried out.

It is the aim of this paper to briefly present the history of the corpus approach in Slovenian linguistic studies and the existing corpora of the Slovenian language, as well as to draw attention to the linguistic studies of the past few years based on this approach. In the second part of the paper, one of the procedures in lexical corpus analysis is presented: the selected examples show the possibility to track changes in the lexis of the Slovenian language in the last decade of the twentieth century by selecting lexical elements introduced into the language with the arrival of the Internet. In addition to showing the dynamics of lexical development, it is our goal to demonstrate the response of the speakers of Slovenian in their acceptance of English lexical elements and their integration into Slovenian.

2 Brief overview of history

Just as the pre-computer corpus SEU, Survey of English Usage, which began in the second half of the 1950s, was a turning point in the linguistic description of English (Kennedy 1998: 19), the collection of materials compiled for the design of Slovar slovenskega knjižnega jezika (1970–1991) (Engl. *Dictionary of the Standard Slovenian Language*), was a turning point for Slovenian lexicosemantic descriptions since it enabled a thorough description of the Slovenian language on the basis of data on textual reality. In the 1960s, when the concept of the new monolingual dictionary was fully formed, lexical descriptions based on materials collected for that purpose, which rejected descriptions of linguistic elements not based on real language use and exceeded the normative approach to language description, were designed.

> Because of the threat to the existence of our nationality, the Slovenians, perhaps more than other nations, are used to being very careful so as not to introduce too many foreign elements or elements not attested to by the literary tradition into our standard language. The dictionary will register much more now: that, which has been recognised as good, less good, or even bad. We tried to show the standard language in its broadest sense of the word: alive, full, with synonyms, inner oppositions, parallel simultaneous norms, a language in its momentum and development. /.../ The dictionary will register the actual state of the language, the bases of its norms, while labels and indicators will be used to show special features, double forms and exceptions (Suhadolnik 1968: 4–5).

About ten years after the first computer corpus, the Brown Corpus, which was created approximately at the same time as the pre-computer corpus for the Slovar slovenskega knjižnega jezika (Engl. *Dictionary of the Standard Slovenian Language*), Croatians began designing their first corpus, based on the American Brown Corpus. Formally, the work began in 1975; the aim of the project was to build a million-word corpus of contemporary Croatian texts (Moguš et. al. 1999: 6). This ambitious project demonstrates the remarkable ability of Croatian linguistics to respond to the trends in American and European linguistics of the time. It is, however, interesting that Sloveni-

an linguistics offered no active response, even though the need for »developing a section dedicated to computational linguistics (with a focus on linguistics)« was recognised at the conference on the Slovene language in Portorož in 1979 (Pogorelec 1983: 113–114). Individual studies, such as T. Korošec's PhD thesis (1976), prove that certain linguists pursued the ideas of an automatic linguistic analysis in Slovenia as well. In the 1980s the field of computer-assisted language data processing began to develop dynamically; proceedings from academic conferences on this topic (*Računalniška obdelava lingvističnih podatkov*, Engl. *Computer Processing of Linguistic Data* 1982, 1985) testify to this, but it remained a peripheral research topic of Slovenian language studies and Slovenian language researchers rarely participated in research on this topic (Korošec et. al 1982). In general, the topic was not explored by Slovenian language researchers and it was computer experts who initiated all the research. It is a pity that Slovenian language studies did not focus on the field of language technology research more, since an excellent opportunity to begin actively developing the field of language technologies of the Slovenian language was missed. This meant that Slovenian language studies only began to focus on language technologies in the second half of the 1990s and started to actively shape this field. Most of the activities were connected with language resource design, especially corpus design.

3 Slovenian language corpora

There are quite a few of corpora available for the Slovenian language; most of them were designed in the second half of the 1990s. The exploration of corpus-building largely began within the framework of an international project, MULTEXT-EAST, which resulted in small literary and newspaper text corpora of Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovenian. In their creation, standards for corpus design and linguistic annotation tools, used earlier in the MULTEXT project, were tested (Erjavec et al. 1995: 88–89). In the second half of the 1990s, the necessity of building larger corpora of the Slovenian language became apparent.

At the moment, there are two monolingual corpora available for the Slovenian language. The first is the 100-million-word reference corpus of the Slovenian language, the FIDA Corpus, a result of co-operation of two research/pedagogical and two commercial partners, Faculty of Arts, University of Ljubljana, Jožef Stefan Institute, DZS Publishing House in Amebis Ltd. The corpus was collected between 1997 and 2000, it is available at http://www.fida.net; Amebis Ltd. also developed concordance software ASP32 (http://www.amebis.si) for corpus analysis of the FIDA Corpus. Unlike the FIDA Corpus, which is a reference corpus, the other, and currently, largest corpus, Nova beseda, a corpus of over 160-million words at the Institute of the Slovenian Language ZRC SAZU has no ambition to be a reference corpus; the largest part of the corpus is composed of texts from Delo, a daily newspaper, (http://bos.zrc-sazu.si/s_beseda.html); but it is currently the largest, freely accessible corpus of the Slovenian language.

At the moment, a new large reference corpus of the Slovenian language, FidaPLUS, (http://www.fidaplus.net) is being created. It is an open-ended corpus to which

texts are constantly being added; the individual segments will gradually become more balanced and it will include a segment of a spoken subcorpus (http://gandalf.aksis.uib.no/tale/ssp/adgang.html). This introduces an entirely new dimension in both quantity and quality of language resource design in the Slovenian context.

	FIDA	**Nova beseda**	**Plans for FidaPLUS**
Type of corpus	synchronic static reference written (the only spoken segment: transcriptions of parliamentary discussions)	synchronic -diachronic dynamic non-reference written (the only spoken segment: transcriptions of parliamentary discussions)	synchronic dynamic reference written + a pilot spoken segment+ a sample of Slovene Internet archive
Format	SGML TEI	special format in the EVA editor/an XML version	XML TEI
Linguistic annotation	automatic lemmatization automatic morphosyntactic tagging	no linguistic annotation	automatic lemmatization automatic morphosyntactic tagging
Tools for analysis	ASP32	Neva	ASP32 and Bonito
Size	100 million	162 million	300 million; 100 million balanced
Accessibility	free access for researchers in the institutions involved in the project, other users are charged a fee	free access	free access for non-commercial use with user registration

Table 1: *Basic data on the type and characteristics of the FIDA Corpus, the Nova beseda Corpus and the FidaPLUS Corpus.*

Ensuring a permanent dynamic growth of a reference corpus will have to be one of the priorities in language resource design for Slovenian in the future, but there is also a growing need to consider the Web as a corpus for Slovenian, with all its limitations, since we need to be aware that the ideas which work with English cannot simply be transferred to Slovenian. The importance of a dynamic reference corpus is well-illustrated by a topical expression referring to a new genre, which has appeared fairly recently in Slovenian, but has quickly become naturalized and can motivate in the sense of word-formation, i.e. *blog*.

FIDA	Nova beseda	Najdi.si
blog bloger	**blog** bloger blogger blogerski bloggerski	**blog** **blogg** blogar blogarica bloger blogerka blogerski blogger bloggerjev bloggec blogati bloganje

Figure 1: *The term* blog *and its derivations in the FIDA Corpus, the Nova beseda Corpus and on the Najdi.si website* [*5 November 2005*].

Parallel corpora for Slovenian exist only in combination with English so far, in spite of the tendency for different language combinations. An English-Slovenian corpus, ELAN, (http://nl.ijs.si/elan) was made within the framework of a European project, the corpus project of students of Translation at the Faculty of Arts, University of Ljubljana, TRANS, http://www-ai.ijs.si/čspela/trans-index.html, is similar to ELAN, while Evrokorpus, http://www.sigov.si/evrokor/, a parallel corpus was produced as an upgrade of the terminological database created in the translation of European legislation.

4 Lexicosemantic corpus descriptions of the Slovenian language

We now leave aside the lexicosemantic descriptions of the Slovenian language based on pre-computer language corpora, above all the Slovar slovenskega knjižnega jezika (Engl. *Dictionary of the Standard Slovenian Language*) (1970–1991) and the lexicosemantic studies based on this dictionary (Vidovič Muha 2000). As mentioned above, they are an extremely important segment in the development of Slovenian linguistic studies which was made possible above all by the data on language reality. We would like to focus on the segment of corpus-based descriptions, i.e. the empirical analysis of samples of language in use as manifested by a corpus with automatic and interactive techniques.

Corpus linguistics has successfully completed its first phase, which is, of course, essential for any further development, with the completed projects of corpus building. The inevitable interdisciplinary approach in corpus design has helped create a solid basis for a broad development of the field. The existing Slovenian language corpora have also provided an incentive for a series of thorough corpus studies, both monolingual and contrastive (Gorjanc 2002, 2005b, Vintar 2003, Gantar 2004, Pisanski Peterlin 2005). At the same time, corpora, especially the FIDA Corpus, are increasingly becoming an indispensable part of language research in general, above all in lexical

and lexicosemantic studies (e.g. Gorjanc and Krek 2001, Jakopin 2001, Vintar 2001, Drstvenšek 2003, Krek 2003, Vintar and Gorjanc 2003, Erjavec and Vintar 2004, Krek 2004, Gorjanc, Krek and Gantar 2005, Holz 2005, Žagar 2005), many of which are also phraseological studies (e.g. Gantar 2003, Kržišnik 2003).

Just as for other languages, the introduction of corpora in language descriptions meant important dictionary projects for Slovenian as well. Unfortunately, corpora have not provided an incentive for monolingual lexicography, but it was in the stage of the design of the new, comprehensive English-Slovenian dictionary that the FIDA Corpus, which later became the basis for the Slovene part of the Oxford-DZS English-Slovenian dictionary (Simon Krek, Ed., 2005: *Veliki angleško-slovenski slovar Oxford*. A–K. Ljubljana: DZS. 1035 pages), began to be created. This is the first dictionary into which the corpus data of Slovenian is incorporated (Grabnar and Šorli 2003).

4.1 An example of a lexicosemantic corpus analysis

To illustrate how structured language data in a corpus can be used for lexical analyses, we present here one of the examples of a lexical corpus analysis of the Slovenian language which is only possible with a large quantity of machine-readable language data. The starting point of the analysis involved comparing the wordlist from the FIDA Corpus with the list of new terms in English, as presented by J. Ayto (1999). By means of corpus analysis, we tried to determine when a lexical element motivated in English occurs in the Slovenian language and how it establishes itself in the language. Since pairs of synonyms or strings often occur with new lexical elements, we tried to determine these relations as well. With the help of markers of semantic relations already identified for the Slovenian language by corpus analysis (Vintar and Gorjanc 2003), we identified pairs of synonyms and strings within the corpus, and studied the dominance of one or the other element in the pair of synonyms.

4.1.1 Obtaining corpus data on pairs of synonyms and strings

Semantically related lexemes often appear in predictable contexts; that is why it is possible to identify semantically connected lexis on the basis of samples of mutual textual connections from the corpus. The starting point was determining the text markers of semantic relations; a corpus analysis based on a subcorpus of natural science and technical texts from the FIDA Corpus and examples from research in other languages (Meyer et al. 1999; Pearson 1998: 174–175) has revealed the following relevant text elements which function as interlexeme semantic relation markers (Vintar in Gorjanc 2000) for Slovenian:

- for synonymy: *ali*, *ali tudi*, *imenujemo (tudi)*, *imenovan tudi*, *sinonim*, *je sinonim za*, *znan tudi kot*, *znan tudi pod imenom, je poimenovan, nosi ime*... (Engl. *or, also, we (also) call it, also called, a synonym, is a synonym for, also known as, also referred to as, is called, is named...)*
- for hyper- and hyponymy: *je*, *kot je (na primer)*, *kot je npr.*, *je vrsta*, *prištevamo*

med, *sodi med*, *med * sodi*, *spada med*, *spada v družino*, *uvrščamo med*, *med * uvrščamo*, *uvrščamo v skupino*... (Engl. *is, such as (for instance), e.g., is a type of, is classified among, belongs to, belongs among, belongs to, belongs to the family, is classified among, is classified in the group*...)

- for meronymy: *ima*, *ima * dele*, *je iz*, *je sestavljen iz*, *vsebuje*... (Engl. *has, has * parts, is made of, is composed of, contains* ...)

Among the above listed markers, the connectors *ali* (Engl. *or*) and *ali tudi* (Engl. *also*) are irrelevant for corpus analyses with the analytical procedures used here, since they cover too many different text functions and yield poor results in terms of identifying two terminological synonyms. The situation is quite different with regard to some other semantic markers, such as *imenovan tudi/imenujemo tudi* (Engl. *also called /we also call it*).

opisan neposreden način odkril dušikov oksid,	**imenovan tudi**	smejalni plin, zaradi katerega postane človek
Vitamin B1,	**imenovan tudi**	tiamin, je verjetno najbolj znan med šestimi vitamini
Vitamin B2,	**imenovan tudi**	riboflavin, je pravzaprav deležen najmanj pozornosti
Stopnjo dostopa do kode	**imenujemo tudi**	doseg procedure.
rumenkastorjave maroge. Ta samotarski kuščar,	**imenovan tudi**	žlezoglavi legvan, je v preteklosti
že kdaj slišal(-a), da Zemljo	**imenujemo tudi**	modri planet?
Zato spletne strani	**imenujemo tudi**	HTML dokumenti. V osnovi je HTML dokument
Večplastno osebnost	**imenujemo tudi**	razcepljena osebnost; to je izraz, s katerim
karte meril 1 : 10 000 in 1 : 5 000	**imenujemo tudi**	detajlne geološke karte, karte v še večjih merilih
Oddajanje hitrih elektronov	**imenujemo tudi**	sevanje žarkov ß, ves pojav pa
Snovi v trdnem agregatnem stanju	**imenujemo tudi**	trdnine. Tudi pri njih nas zanima, kako se

Figure 2: *Part of the concordance string for the search condition* imenovan tudi/imenujemo tudi (Engl. *also called /we also call it*).

The marker of synonymy *imenujemo tudi* (Engl. *we also call it*) actually shows true synonyms, e.g. *dušikov oksid – smejalni plin* (Engl. *nitrous oxide – laughing gas)*, *vitamin B1 – tiamin* (Engl. *vitamin B1 – tiamine)*, *vitamin B2 – riboflavin* (Engl. *vitamin B2 – riboflavin)*, *dostop do kode – doseg procedure* (Engl. *code access – procedure scope*), *spletna stran – HTML dokument* (Engl. *web page – HTML document)*. At the same time, it turns out that it connects not only lexical synonyms, but also the lexeme and its paraphrase, e.g. *Trdine so snovi v trdnem agregatnem stanju*, *Železnata tla so tla, bogata predvsem z železovimi spojinami* (Engl. *Solids are materials in the solid phase, Ferrous soil is rich above all in iron compounds*) etc.

Punctuation marks in their non-syntactic role, above all quotation marks and parentheses, also occur as interlexeme relation markers; they generally mark pairs of synonyms by including the synonym which is less frequent, uncommon or foreign in origin (Gorjanc 1996: 256–257). It is also possible to obtain information on synonyms from a corpus by using these two types of punctuation marks, but it has turned out that as the parentheses above all, are multifunctional, the analyses fail to yield relevant results. However, if we limit the search to a specific part of the corpus, e.g. natural science texts (Cobiss label *Natural sciences*), and to adjacent noun + noun combinations, the results are encouraging.

enocelični plazmodiji razgrajajo rdeča krvna	**telesca (eritrocite)**	in ob tem povzročajo silne napad
v vodik in kisik. Vodik se nabira na negativni	**elektrodi (katodi)**	, kisik pa na pozitivni
lastnosti dimnih zaves temeljijo na optičnih pojavih	**disperzije (razprševanja)**	in absorpcije (vsrkanja) svetlobe
dneh na zemeljski ekvator (polutnik) ter na oba	**pola (tečaja)**	, severnega in južnega. Če naprej
tega ima sodobna kopija kar 8-krat večji delovni	**pomnilnik (RAM)**	In 4-krat večji trajni pomnilnik (ROM)
kemijski postopek, kako iz slanice pridobivati natrijev	**hidroksid (lug)**	, ki je za izdelavo mila neprimerno boljši
lastnosti sta hitro učinkovanje in visoka stopnja	**strupenosti (toksičnosti)**	; so brez barve, vonja in okusa.
dela ali telesa nevrona, več krajših, vejastih	**izrastkov (dendritov)**	in le enega dolgega izrastka (aksona).
Je pri svojih operacijah uporabljal karbolno	**kislino (fenol)**	, da je preprečil zastrupitve. Kasneje so
sušijo, potem ko so jih prepojili s polietilen	**glikolom (PEG)**,	v vodi topljivo polimerno smolo, katere
sestava je odvisna od matične kamnine, odnašanja	**prsti (erozije)**	in živih bitij, ki sodelujejo pri nastajanju
Ptiče bogov in kraljev, ki se v času	**ženitve (spomladi)**	v resnici prelevijo v pravljična bitja.

Figure 3: *Part of an edited concordance string for the search* Noun (Noun) *in the subcorpus »natural sciences« (Cobiss).*

Once the concordance string is edited and only pairs or synonyms are left, it turns out that parentheses as markers of synonymy generally occur with lexicalised semantic pairs, e.g. *rdeče krvno telesce – eritrocit* (Engl. *red blood cell – erythrocyte*), *karbolna kislina – fenol* (Engl. *carbolic acid – phenol*), *odnašanje prsti – erozija* (Engl. *soil loss – erosion*), etc., while pairs of synonyms where a text actualisation is used as a synonym are rare, e.g. *čas ženitve – spomladi* (Engl. *time of marriage – in the spring*). The text sample is thus effective for obtaining pairs of synonyms from the text; the pairs of synonyms are above all of the type loan word – Slovenian word or acronym – phrase.

4.1.2 The distribution of selected pairs of synonyms or concordance strings in the FIDA Corpus

It is possible to follow the relations between pairs of synonyms and synonym strings with the aid of corpus data. Corpus data will reveal the dominant term in a pair of synonyms or a string, and, according to the information on time distribution, the change in the dominant term with usage preference in a discourse community.

Corpus data can thus be used to realise the principle of synchrony, based on European structuralism. Due to the nature of language data, synchrony has often been equated with synchronic statics; this, however, was not the original idea of structuralism:

> It would be a serious mistake to consider statics and synchrony to be synonyms. Static section is a fiction: it is not a special form of scientific procedure, only its auxiliary method. The perception of a film may be considered not only diachronically, but also synchronically: however, the synchronic view of a film is not identical with an isolated picture extracted from the film. The perception of movement is present even in synchronic view. The same is true of language (Jakobson 1931: 264–265). (English translation form: Dictionary of the Prague school of linguistics. (Ed.) Libuše Dušková. Amsterdam, Philadelphia: John Benjamins, 2003, p.154)

Dynamic corpora above all, to which new texts are continuously added, are truly able to follow the development of a language; at the same time they reflect decisions

of the discourse community. This can be seen from the example of an analysis of the lexical element *(svetovni) splet* (Engl. *(World Wide) Web*) entering the Slovene universe of discourse in the second half of the last decade.

In the two years after its first appearance, only the loan word occurs in the corpus, but when the Slovenian variant appears, it immediately becomes a successful rival and the use of the loan word gradually decreases (Gorjanc 2005b: 115).

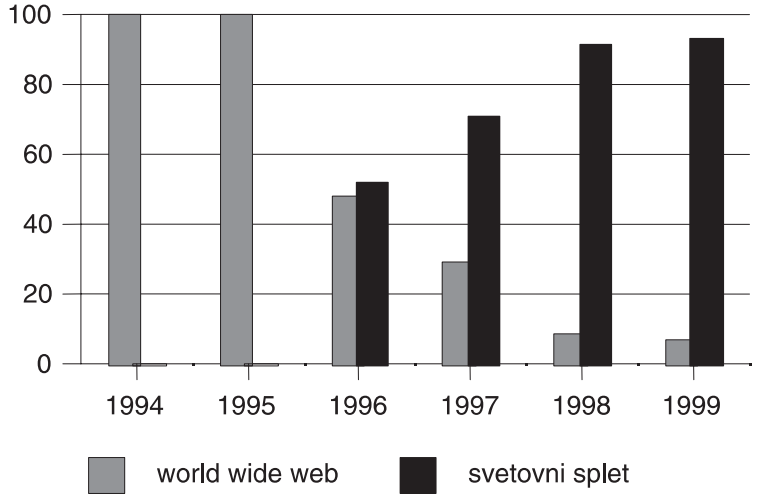

Figure 4: *Proportion of terms for* WWW *between the years 1994 and 1999 in the FIDA Corpus.*

In written texts, the dominance of the Slovenian synonym over the loan word is even more obvious in the case of another key term from the field of the Internet, i.e. *home page*. After eliminating corpus noise related to proper names of pages, it turns out that the Slovenian term *domača stran* (Engl. *home page*), has dominated completely (91.8 % of corpus occurrences). In addition to the calque *domača stran* (Engl. *home page*), there is also a rival new term *predstavitvena stran* (Engl. *presentation page*) (6.8 %), but it seems that the motivation in the calque from English is more acceptable. The opposite occurs with the term *screen saver.*

In addition to the loan word, the calque *varčevalnik zaslona* occurs next, but a later Slovenian term formed by using the attribute *ohrajeva-* (Engl. *keep*) turns out to be more acceptable. Two derivational variants occur, but later the derivative from the adjective with the suffix *-ik* dominates.

The term *internet* itself is now fully integrated in the Slovenian language; this is partly due to its everyday use. As a noun, it occurs as a premodifier in noun phrases: e.g. *internet storitev* (Engl. *Internet service*), *internet naslov* (Engl. *URL*), *internet povezava* (Engl. *Internet connection*), *internet ponudnik* (Engl. *Internet service provider*), *internet stran* (Engl. *Web page*), *internet račun* (Engl. *Internet account*), *internet protokol* (Engl. *Internet protocol*). The noun *internet* happens to be extremely prolific in terms of word formation, since it forms:

- derived classifying adjectives ending in *-ni*, and *-ski*: *internetni*, *internetski*,
- a derived classifying adjective ending in *-ov*: *internetov*
- a classifying adjective of a higher degree of derivation in *-ski* : *internetovski*,;
- an adverb derived from the classifying adjective in *-ski*: *internetsko*
- a noun derived from a noun and a noun of a higher degree of derivation derived from an adjective: *internetar*; *internetovec*, as well as
- a compound noun meaning »internet addict« *internetdžanki* (Engl. *Internet junkie*).

In classifying adjectives, the variability is relatively high, that is why we attempted

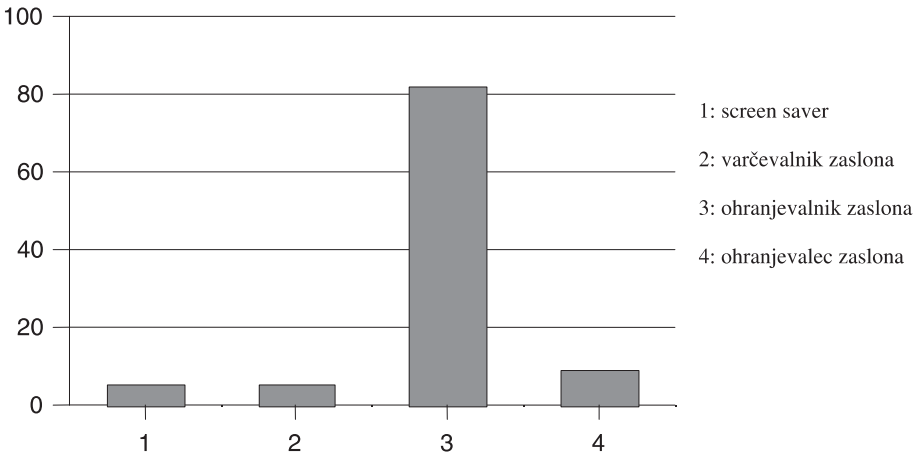

Figure 5: *Relations in the synonym string for 'screen saver' in the FIDA Corpus.*

to determine whether the corpus can reveal information on links between an individual variant and specific strings of co-occurrences. It turns out that the collocators of the adjectives *internetni*, *internetski* and *internetovski* overlap /service, page, search engine, business, shop, bookseller, service provider.../, so that it is impossible to determine the specific phrases in the individual instances. Therefore, it seems that the use is very much optional and different variants of the adjective are possible with the same headword. In the case of the adjective *internetov*, which is the least common of the adjectives listed above, the link to the headword is completely dispersed; this indicates that the suffix variant *-ov* is not integrated and consequently inappropriate for the classifying character of the adjective. The frequent use of the classifying adjective with the suffix *-ni* (*internetni*) shows a prevalence of this variant, its only real rival is the classifying adjective with the suffix *-ski* (*internetski*).

The corpus analysis in the FIDA Corpus for another pair of synonyms, *internet – medmrežje* (Engl. *the Internet*), with the search conditions for *internet** and *medmrež** , yields the ratio 13,638 : 308; at the same time we find that *medmrežje* is not productive in terms of word formation. This confirms the fact that the attempt to coin a new term was unsuccessful, although the *Slovar slovenskega pravopisa* (2001) (Engl. *Slovenian orthographic dictionary*) prescribes *medmrežje* as the more acceptable synonym in the pair of synonyms referring to the Internet.

5 Conclusion

In the last decade, corpus linguistics has been very influential in the Slovenian linguistic community. The initial stage was the design of Slovenian language corpora; this was a necessary condition for a further development of the field. Since 2000 the first thorough studies in corpus linguistics have been carried out. Corpora are increasingly becoming the bases of linguistic analyses as an independent research starting point, while at the same time they present the basic research material in various types of linguistic studies. The language data found in a corpus is practically unlimited, and its analysis is a permanent challenge, above all when it surpasses the limits of the expected and breaks our intuitive assumptions about the linguistic reality. The results of corpus analyses of the Slovenian language are exciting; they reveal the great creativity and vitality of the Slovenian discourse community.

V angleščino prevedla
Agnes Pisanski Peterlin.

References

John Ayto, 1999: *20th Century Words*. Oxford: Oxford University Press.

Douglas Biber, Susan Conrad in Randi Reppen, 1998: *Corpus Linguistics. Investigating Language Structure in Use*. Cambridge: Cambridge University Press.

František Čermák, 2002: Today's corpus linguistics. Some open questions. *International journal of corpus linguistics* 7/2. 265–282.

Nina Drstvenšek, 2003: Vloga besedilnega korpusa pri postavitvi geselskega članka v enojezičnem slovarju. *Jezik in slovstvo* 48/5. 65–81.

Tomaž Erjavec, Nancy Ide, Vladimír Petkevič in Jean VÉronis, 1995: MULTEXT-EAST: Multi-lingual Text Tools and Corpora for Central and Eastern European languages. Heike Retting s sodelovanjem Júlie Pajzs in Gáborja Kissa (ur.): *TELRI: »Language Resources for Language Technology«*. Proceedings of the First European Semminar, Tihany, September 15–16. 87–97.

Tomaž Erjavec in Špela Vintar, 2004: Korpus kot podpora slovarju informacijskega izrazja slovenskega jezika. *Uporabna informatika* 12/2. 97–106.

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

– – 2004: *Frazem in njegovo besedilno* okolje. Doktorska disertacija. Mentorica A. Vidovič Muha. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.

Vojko Gorjanc, 1996: Terminologija novejših naravoslovno-tehničnih strok (Ob primeru računalništva in jedrske fizike). Ada Vidovič Muha (ed.): Jezik in čas. Ljubljana: Znanstveni inštitut Filozofske fakultete. 251–260.

– – 2002: Jezikoslovna načela gradnje računalniških besedilnih zbirk strokovnih jezikov. Doktorska disertacija. Mentorica A. Vidovič Muha. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.

– – 2003: Odkrivanje leksikalnih sprememb s pomočjo korpusa. Stanisław Gajda in Ada Vidovič Muha (ed.): *Współczesna polska i słoweńska sytuacja językowa*. Opole: Uniwersytet

Opolski, Instytut Filologii Polskiej/Ljubljana: Univerza v Ljubljani, Filozofska fakulteta. 99–111.

– – 2005a: Tracking lexical changes in the reference corpus of Slovene text. *Corpus Linguistics Around the World*. Amsterdam/New York: Rodopi. 91–100. V tisku.

– – 2005b: *Uvod v korpusno jezikoslovje*. Domžale: Izolit.

– – 2005c: V mavrici jezikovnih podatkov. Vojko Gorjanc in Simon Krek (ed.): *Študije o korpusnem jezikoslovju*. Ljubljana: Krtina. 173–199.

Vojko Gorjanc in Simon Krek, 2001: A corpus-based dictionary database as the source for compiling Slovene-X dictionaries. *Proceedings of the COMPLEX 2001 6th Conference on Computational Lexicography and Corpus Research*. 41–47.

Vojko Gorjanc, Simon Krek in Polona Gantar, 2005: Slovenska leksikalna podatkovna zbirka. *Jezik in slovstvo* 50/2. 3–19.

Katarina Grabnar in Mojca Šorli, 2003: Novi veliki angleško-slovenski slovar Oxford-DZS. Jezik in slovstvo 48/3-4. 126–133.

Nanika Holz, 2005: Mesto *Velikega slovarja tujk* v slovenski leksikografiji. *Jezik in slovstvo*, letnik 50/1. 87–99.

Roman Jakobson, 1931: Prinzipen der historischen Phonologie. *Travaux du Cercle Linguistique de Prague* 4. Prague 1929–1939. 247–267.

Primož Jakopin, 2001: Words and nonwords as basic units of a newspaper text corpus. *Proceedings of the COMPLEX 2001 6th Conference on Computational Lexicography and Corpus Research*. 49–65.

Graeme Kennedy, 1998: *An Introduction to Corpus Linguistics*. London: Longman.

Tomo Korošec, 1976: Poglavja iz strukturalne analize slovenskega časopisnega stila. Doktorska disertacija. Mentor J. Toporišič. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.

Tomo Korošec, Denis Poniž, Peter Tancig, 1982: Uporabnost računalniških konkordanc v lingvističnih in literarnih raziskavah. *Zbornik II. znanstvenega srečanja Računalniška obdelava lingvističnih podatkov*. Ljubljana: Institut Jožef Stefan. 405–415.

Simon Krek, 2003. Sodobna dvojezična leksikografija. *Jezik in slovstvo* 48/1. 45–60.

– – 2004: Slovarji serije COBUILD in formalizacija definicijskega jezika. *Jezik in slovstvo* 49/2. 3–16.

– – (ur.): *Veliki angleško-slovenski slovar Oxford*. 1. knjiga. A–K. Ljubljana: DZS.

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

Milan Moguš, Maja Bratanić in Marko Tadić, 1999: *Hrvatski čestotni riječnik*. Zagreb: Zavod za lingvistiku Filozofskog fakulteta Sveučilišta u Zagrebu & Školska knjiga.

Ingrid Meyer, Kristen Mackintosh, Caroline Barriere in Tricia Morgan, 1999: Conceptual sampling for terminological corpus analysis. Peter Sandrini (ed.): *Proceedings of TKE '99*. Vienna: TermNet. 256–267.

Jennifer Pearson, 1998: *Terms in Context*. Amsterdam: John Benjamins.

Agnes Pisanski Peterlin, 2005: *Konvencije rabe medbesedilnih elementov.* Doktorska disertacija. Mentorica I. Kovačič. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.

Breda Pogorelec (adapted), 1983: Slovenski knjižni jezik, zgodovina slovenskega knjižnega jezika in stilistika. *Slovenščina v javnosti. Posvetovanje o jeziku. Portorož, 14. in 15. maja 1979. Gradivo in sporočila*. Ljubljana: Republiška konferenca SZDL Slovenije in Slavistično društvo Slovenije. 110–114.

Stane Suhadolnik, 1968: Koncept novega slovarja slovenskega knjižnega jezika. *4. seminar slovenskega jezika, literature in kulture. Predavanja iz jezika*. 1–11.

Wolfgang Teubert, 1999: Korpuslinguistik und Lexikographie. *Deutsche Sprache* 99/4. 292–313.
Ada Vidovič Muha, 2000: *Slovensko leksikalno pomenoslovje*. *Govorica slovarja*. Ljubljana: Znanstveni inštitut Filozofske fakultete.
Špela Vintar, 2001: Using parallel corpora for translation-oriented term extraction. *Babel* 47/2. 121–132.
– – 2003: *Uporaba vzporednih korpusov za računalniško podprto ustvarjanje dvojezičnih terminoloških virov.* Doktorska disertacija. Mentor R. Šušteršič. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.
Špela Vintar in Vojko Gorjanc, 2003: Identifying markers of semantic relations in Slovene. *Strani jezici* 1–2. 37–44.
Mojca Žagar, 2005: Determinologizacija (na primeru terminologije fizike). *Jezik in slovstvo* 50/2. 35–48.

Slovene language corpora
Beseda http://bos.zrc-sazu.si/main_si_l2.html [5. 11. 2005]
ELAN http://nl.ijs.si/elan [20. 9. 2005]
Evrokorpus http://www.sigov.si/evrokor/ [20. 9. 2005]
Korpus slovenskega jezika FIDA http://www.fida.net [20. 9. 2005]
Korpus slovenskega jezika FidaPLUS http://www.fidaplus.net [20. 9. 2005]
Multext-East http://nl.ijs [20. 9. 2005] Nova beseda http://bos.zrc-sazu.si/s_beseda.html [20. 9. 2005]
TALE korpus – pilotni govorni korpus slovenskega jezika http://gandalf.aksis.uib.no/tale/ssp/adgang.html [5. 11. 2005]
TRANS http://www-ai.ijs.si/~spela/trans-index.html [20. 9. 2005]

Povzetek

Korpusno jezikoslovje se je v zadnjem desetletju dokončno uveljavilo kot posebno raziskovalno izhodišče, utemeljeno strogo empirično, v zadnjih nekaj letih tudi v slovenskem prostoru kot ločeno raziskovalno izhodišče. Nujni predpogoj za to so bili seveda korpusi, zato je druga polovica devetdesetih let prejšnjega stoletja zaznamovana z njihovo gradnjo, pri čemer so pionirsko vlogo odigrali korpusi, nastali v okviru mednarodnega projekta MULTEXT-EAST. Danes imamo za slovenščino na voljo dva enojezična korpusa, 100-milijonski referenčni Korpus slovenskega jezika FIDA, ter večji, a nereferenčni Nova beseda, velikosti nekaj nad 160 milijonov besed; v izgradnji pa je obsežni 300-milijonski referenčni korpus FidaPLUS. Ob tem so bili oblikovani tudi vzporedni korpusi, zaenkrat samo v jezikovnem paru z angleščino. Tako oblikovani korpusi so osnova za vrsto korpusno utemeljenih jezikoslovnih študij, nastalih v zadnjih letih. Kot je za angleški prostor pomenila veliko prelomnico pri jezikovnih opisih predračunalniška besedilna zbirka Survey of English Usage, je bila to za slovenske leksikalnopomenske opise predračunalniška gradivna zbirka, nastala za potrebe izdelave Slovarja slovenskega knjižnega jezika (1970−1991), saj je omogočila celovit leksikalni opis slovenskega jezika na podlagi podatkov o besedilni realnosti. Ko se je v šestdesetih letih prejšnjega stoletja dokončno oblikoval koncept novega enojezičnega slovarja, so se v slovenskem prostoru načrtovali leksikalni opisi, temelječi na obsežnem gradivu, ki so zavračali možnost opisa jezikovnih elementov brez podlage v jezikovni realnosti in presegali normativistični pristop k jezikovnemu opisovanju. Kljub takemu programskemu izhodišču pa v tem času v okviru slovenistike ni prišlo do oblikovanja računalniško podprtega dela z jezikovnimi podatki, čeprav je bilo to eno od njenih ekspliciranih programskih izhodišč. Tako se je slovenistika zares priključila

oblikovanju področje jezikovnih tehnologij za slovenski jezik šele v drugi polovici devetdesetih let prejšnjega stoletja, vendar takrat zelo opazno, tako da lahko ugotovimo, da je korpusno jezikoslovje je v zadnjem desetletju pomembno zaznamovalo slovenski jezikoslovni prostor, še posebej po letu 2000, ko na osnovi oblikovanih korpusov dobimo prve celovite korpusnojezikoslovne študije. V slovenistiki so korpusi postali po eni strani izhodišče jezikovne analize kot samostojnega raziskovalnega izhodišča, po drugi pa so v različnih tipih jezikoslovnih raziskav nujno potrebni kot gradivna osnova jezikoslovnega raziskovanja. Korpusni jezikovni podatki so praktično brezmejni, njihova analiza nenehen izziv, še posebej takrat, ko presegajo meje pričakovanega in rušijo naše intuitivne predstave o jezikovni realnosti. Rezultati korpusnih analiz slovenskega jezika so v veliki meri navdušujoči; razkrivajo namreč izjemno kreativnost in vitalnost slovenske diskurzivne skupnosti.