English Anchors in a Slovenian Word Resource

Primož Jakopin[1] and Andreja Žele[2]

1. Introduction

Word resources remain as ever one of the fundamental tools not only for the linguistic research but also for the now slowly maturing related fields such as machine translation, natural language processing or speech recognition. For the languages with smaller number of speakers such as Slovenian (approx. 2 million), where competing commercial projects would be difficult to justify, basic lexicographic work has traditionally been the domain of the academia. Fran Ramovš Institute of the Slovenian Language has in the early nineties of the past century completed two monolingual resources, the *Dictionary of Standard Slovenian* (SSKJ, 93.000 entries) and the *Dictionary of Lesser Used Slovenian Words* (BSJ, 178.000 entries). The latter contains the words which have been given close attention by the SSKJ team for the possible inclusion in the main dictionary but did not make it through. As opposed to SSKJ which is a regular monolingual dictionary with all the relevant headword data, the entries in BSJ contain only accent, part-of-speech and basic inflectional data. Further context, compiled during the preparation of SSKJ, is not available in digital format, and many entries are not based on typewritten, but only on handwritten evidence.

Table 1: 18 entries from the Dictionary of Lesser Used Slovenian Words

abstrakcíjski, -a -o, prid.	*abstraktizíranje, -a, s
abstrakcioníst, -a, m	abstraktizírati, -am, nedov. in dov.
*abstrakcionístičen, -čna -o, prid.	*abstráktnohumanístičen, -čna -o, prid.
abstrakcionízem, -zma, m	abstráktor, -ja, m
*abstrakcízem, -zma, m	*abstrúz, -a, m
*abstraksíst, -a, m	absurdíst, -a, m
*abstráktičen, -čna -o, prid.	absurdístičen, -čna -o, prid.
*abstraktivíranje, -a, s	absurditéta, -e, ž
*abstraktivízem, -zma, m	absurdízem, -zma, m

To assist the further dictionary preparation, there are several such projects ongoing, including a *Dictionary of Recent Vocabulary in Literary Slovenian* (Žele 2005), a text corpus *Nova beseda* or *New word* in English, has been in operation at the Institute since 2000 (Jakopin 2001 b). It now contains 202 million words, and it is available at the URL http://bos.zrc-sazu.si/a_beseda.html. A natural extension of all this work has

[1] Fran Ramovš Institute of the Slovenian Language ZRC SAZU, Ljubljana, Slovenia
e-mail: primoz.jakopin@guest.arnes.si
[2] Fran Ramovš Institute of the Slovenian Language ZRC SAZU, Ljubljana, Slovenia
e-mail: andrejaz@zrc-sazu.si

been a two year research project (2004–2006), which established a large *Online List of Slovenian Words* (LSW), available at the URL http://bos.zrc-sazu.si/besede_en.html. It contains 356.000 entries from the SSKJ, the BSJ, the *Nova beseda* corpus and from the index of the *NAJDI.SI* Slovenian web search service. Every entry has at least basic inflectional information with part of speech, followed by search links to further Internet resources such as *SSKJ*, *Nova beseda, NAJDI.SI*, Slovenian *Wikipedia*, and, in some instances, an example from the web page. SSKJ entries also have an abridged explanation. Links are present only where relevant – a great bonus here has been the index of the *NAJDI.SI* Slovenian web search engine (over 5 billion word tokens in all), obtained through cooperation with the service provider, Interseek. Though Internet content is volatile, pages keep changing on the fly, it still gives a useful approximation on what is available and what is not. A considerable effort, invested to obtain the search links to the other Slovenian text corpus FIDA plus (http://www.fidaplus.net, 621 million words) has not produced some usable result yet.

Table 2: List of Slovenian Words - search result for the string *korpus* (16 hits)

kórpus	**1.** *vojaška enota iz divizij in posebnih enot:*; m, source: S; links: *sskj*, *najdi.si* *wiki*
kórpusen	*nanašajoč se na korpus 1:*; prid., source: S; links: *sskj*, *najdi.si*
korpúskel	-kla; m, source: B; links:
kórpuski	-a -o; prid., source: B; links:
korpúskul	fiz. *osnovni delec*; m, source: S; links: *sskj*, *najdi.si*
korpúskula	-e; ž, source: B; links: *najdi.si*.
korpuskuláren	*nanašajoč se na korpuskul:*; prid., source: S; links: *sskj*, *najdi.si*
korpuskulárnost	-i; ž, source: B; links:
korpus delikti	source: S; links: *sskj*, *najdi.si*
antikorpus	m, source: I; links: *najdi.si*
córpus delícti	jur. *predmet, ki izpričuje storitev kaznivega dejanja:*; m, source: S; links: *sskj*, *nova beseda*, *najdi.si*
ekstrakorpuskularen	prid., source: I; links: *najdi.si*
nekorpuskularen	prid., source: I; links: *najdi.si*
antrum	m, source: I; links: *najdi.si* Example: *Znano je, da po izkoreninjenju okužbe počasi izzvenevajo vnetne spremembe sluznice v korpusu in antrumu želodca.*
deiksis	m, source: I; links: *najdi.si* Example: *Na podlagi pilotne analize otroškega govora bomo ob obravnavi osebnih deiksisov, to je izrazov s kazalno vlogo, skušali pokazati razliko med govorjenimi in zapisanimi besedili in s tem utemeljiti potrebo po načrtovanem zbiranju govorjenih besedil in njihovo vključitev v korpus slovenskega jezika.*
knójevec	pog., med narodnoosvobodilnim bojem in prva leta po 1945 *član Korpusa narodne obrambe Jugoslavije:*; m, source: S; links: *sskj*, *najdi.si*

As can be seen from Table 2, and even more from the Table 1 (asterisk marked entries) there are quite a few entries in the LSW where no web presence could be established. Examples are the words *abstrakcionístičen*, *abstrakcízem*, *abstraksíst*, *abstráktičen*, *abstraktivíranje,* *abstraktivízem*, *abstraktizíranje*, *abstráktnohumanístičen* and *abstrúz* from Table 1 as well as *korpúskel*, *kórpuski* and *korpuskulárnost* from Table 2. Altogether there are about 81,000 such entries or 45% of the BSJ.

2. English connection

The problem with links in the *List of Slovenian Words* is twofold. On one hand there are entries with no links, which would require the complex digitalisation of the old SSKJ repository, accumulated from the times as far back as before the second world war, and on the other hand there are entries with search links that produce few or inadequate hits. The obvious solution would be to look elsewhere, to the knowledge treasuries in other languages, most notably in English. As the colonisation of Northern America did not proceed under the wing of Slovenian naval forces, the language of the main Internet search engine today, the Google, is different, though from the same wider language domain. Some playing around with the asterisk-marked entries from the Table 1 gives their, mostly English counterparts which all produce some echo, as shown in Table 3. The translations of three entries, *abstraktivíranje*, *abstraktizíranje* and *abstráktnohumanístičen* have a very marginal web presence while the others fare better, with *abstractism* being a non-rare word.

Table 3: Translations of no-link words from Table 1 with Google hit numbers

abstrakcionístičen - abstractionistic, 239
abstrakcízem - abstractism, 17.200
abstraksíst - abstractist, 1.210
abstráktičen - abstractic, 487
abstraktivíranje - Abstraktivierung, 1
abstraktivízem - abstractivism, 927
abstraktizíranje - abstractising, 5
abstráktnohumanístičen - abstractly humanistic, 4
abstrúz - abstruz, 1.210

The most frequent words in Slovenian and English parlance of course have quite a different spelling but quite a lot of words in specialized vocabularies, such as the words in Table 3, have much in common. LSW collection has 356.000 entries so the likelihood of finding suitable translation equivalents with machine help is worth a try.

The best Slovenian-English online resource, provided by Amebis from Kamnik at the foot of the Alps, where 200 characters can be translated at a time in demo mode, is available at the URL: http://presis.amebis.si/prevajanje/. It is based on a dictionary of approx. 50.000 common words. That leaves 306.000 other words, candidates for inclusion in the current research project *Contemporary slovene lexicon* (*online language resources*) or *Novejša slovenska leksika v povezavi s spletnimi jezikovnimi viri* in Slovenian, where both authors are participating.

3. The algorithm

One of the most obvious paths to follow would be to observe the complex words in Slovenian – derivatives with prefix or prefixal formant; among compounds in Slovenian subordinate type is typical, coordinate compounds occur rarely. The first parts of interfixal-suffixal or only interfixal compounds, *prvi deli zloženk* in Slovenian, are a standard topic of linguistic research (see Vidovič-Muha 1988 or Toporišič 2000, p. 193 for Slovenian and Borror 1960 or Sheehan 2000 for English). There is also a simple criterion available for how to divide the potential prefixal derivative or compound into its components – if the second part of the complex word (of derivative or compound) is also a valid entry in the dictionary, and longer than 2 letters, chances that the word actually is a compound and that the splitting point is accurate are considerable. Such an algorithm, when applied to LSW, gives a list of 66.708 possible prefixes with a cumulative frequency of 294.328. Yet the beginning of the list, shown in Table 4, does not look particularly promising.

Table 4: First 20 entries in the prefix-candidate list

a	13364	882	0.07	abaka	1	1	1.00	abderit	5	1	0.20	abdu	6	1	0.17
aa	8	3	0.38	abal	3	1	0.33	abdi	5	1	0.20	abduci	2	1	0.50
aaa	2	1	0.50	abamek	1	1	1.00	abdicira	2	1	0.50	abdukti	1	1	1.00
ab	390	44	0.11	aban	2	2	1.00	abdomino	2	1	0.50	abe	40	8	0.20
aba	18	4	0.22	abde	6	1	0.17	abdominoperinea	1	1	1.00	abec	17	2	0.12

The second column contains the number of occurences of the string as part of beginning of the word, the third one the number of such cases where the second part of the word is also an entry in the dictionary and the fourth the ratio between the 2 values, which can be between close to 0.00 (worst case) and 1.00 (best case - all compounded words with this prefix have a second part which is itself a word in the dictionary). After a descending sort on the ratio the Table 5 contents moves slightly in the direction of what would better suit an empiric mind:

Table 5: Top 20 entries in the prefix-candidate list, sorted on ratio

abaka	1	1	1.00	abei	1	1	1.00	ablak	1	1	1.00	aboliti	1	1	1.00
abamek	1	1	1.00	abek	1	1	1.00	ablas	1	1	1.00	abomasopek	1	1	1.00
aban	2	2	1.00	aberantn	1	1	1.00	abluto	1	1	1.00	abomi	1	1	1.00
abdominoperinea	1	1	1.00	abhi	1	1	1.00	abnormi	1	1	1.00	abondan	1	1	1.00
abdukti	1	1	1.00	abhy	1	1	1.00	abnormn	1	1	1.00	abonma	2	2	1.00

Prefixes with very low total occurrence are on top of the list, as it is abc-sorted inside ratio. Best prefixes tend to be those with high ratio and high valid second compound frequency (column 3). After some experimenting the square of the ratio, multiplied by second compound frequency, used in the role of the sorting key, yields the final list:

Table 6: Top 18 entries of Slov.prefix-candidate list, sorted on ratio derivative (col.4):

Ne	15444	9744	3867.39	za	9016	4568	1188.14	pod	2652	1736	733.46
samo	2808	2360	1665.22	nad	1812	1451	928.64	od	3923	2249	730.70
pred	2774	2208	1413.12	mikro	1083	1018	899.50	super	879	801	663.31
proti	1883	1695	1372.95	brez	1576	1256	803.84	bio	937	830	657.44
Pre	12011	5653	1248.75	pri	4815	2613	761.95	avto	1049	860	578.26
anti	2148	1786	1230.38	vele	890	834	736.92	foto	716	664	574.29

The last three items: *bio*, *avto* and *foto* are not prefixes in the narrow sense but prefixal parts of a compound. A close look at the first 525 entries in the list reveals very little noise, only 5 prefix candidates that do not fit quite well in the picture: *v*, *ra*, *p*, *š* and *i*.It is also interesting to note that out of 369 words, included in the SSKJ (the Dictionary of Standard Slovenian) and labeled as *first part of compounds*, *prvi del zloženk*, just 66 are missing among the first 525 entries of the list from Table 6, and only 10 among the first 5.000.

4. Application to BNC wordlist

For some time now the wordlist from the British National Corpus, with frequencies and pos tags is also available on Internet. The author downloaded it from the site of the former Information Technology Research Institute, University of Brighton (now Natural Language Technology Group), at the URL ftp://ftp.itri.bton.ac.uk/bnc/. It contains 938.972 different tokens with a cumulative frequency of 100.104.253, i. e. word candidates which of course not all are words in the narrow sense (see for instance Jakopin 2001 a). There are 561.266 tokens, composed only of letters, with a cumulative frequency of 95.270.944 (95% of the corpus). The algorithm described in chapter 3, applied to the letter-only tokens and excluding 12 single-letter prefix candidates, gave a list of 73.467 would-be prefixes, some of which are shown below.

Table 7: Top 100 entries in the BNC prefix-candidate list (as prefixes of derivatives or first parts of compounds):

un	de	bio	euro	house	post	land	war	sur	key
re	sub	wood	white	work	pre	home	horse	com	mono
over	back	head	foot	free	night	sky	moon	al	fair
under	multi	down	en	trans	bed	wind	psycho	star	turn
dis	up	photo	tele	sand	side	im	eye	electro	dun
out	water	green	fore	sea	book	mc	neuro	uni	short
inter	air	counter	hand	hyper	immuno	snow	time	hay	radio
micro	black	anti	fire	high	cross	data	bur	body	stock
super	non	auto	play	power	broad	red	video	poly	pro
mis	sun	in	long	con	news	hard	good	ultra	road

The table speaks for itself; the BNC wordlist is otherwise used during different phases of the project where manual checkup is required, for quick lookups.

5. Solution to the problem

Quite a few of the prefixes or first parts of compounds from the algorithmically obtained prefix candidates (see Table 6) is easily and directly translatable to English (*anti*, *mikro*, *super*, *bio*, *avto*, *foto* …). But what about the second parts of such compounds? It turns out that the situation here is not bad either. For the 520 manually checked prefixes (of derivatives) or as first parts of compounds there are 43.795 different second parts with a cumulative frequency of 120.525. And the top 6.512 (down to and including the frequency of 5) cover 60.212 complex words from the List of Slovenian Words. So an outcome, at least comparable in size to the number of available translations (50.000) can be predicted, and with considerably less effort than would be required by a manual translation.. The top 10 such second parts of compounds are shown in Table 8. The first column contains the second component, followed by an approximate English translation where feasible, its frequency in the discovered compounded words and the total frequency – the number of all the words in LSW which end with the given string. Especially the last column shows the full potential of quantitative approach: the pair *-logija* (*-logy*) and *–log* (*-logist*) alone acoounts for over 1.000 occurences.

Table 8: Top 10 second parts of juxtaposed compounds with valid & total frequencies

-kulturen ~ -cultural	91	144
-političen ~ -political	91	160
-zgodovinski ~ -historic(al)	72	105
-tehničen ~ -tehnical	71	153
-logija ~ -logy	68	760
-meter ~ -meter	66	373
-vrsten ~ -*serial*	64	111
-ličen ~ -*neat*	63	599
-log ~ -logist	63	527
-loški ~ -logical	63	870

6. Conclusion

As in any field of linguistic, or more specifically lexical endeavour, the territory is not conquered until the infantry has set foot on it. Yet any tool that shortens the trail is welcome and the method described in the paper will, so we are convinced, make the emergence of English equivalents in the large Slovenian word resource less distant and more feasible.

References

Borror, D. J. 1960. *Dictionary of Word Roots and Combining Forms*. Mayfield, Palo Alto, 134 pp.

Jakopin, P. 2001 a. Words and nonwords as basic units of a newspaper text corpus. In: *6th Conference on Computational Lexicography and Corpus Research "Computational Lexicography and New EU Languages"*, Birmingham, 28 June-1 July, 2001. COMPLEX 2001. Birmingham: Centre for Corpus Linguistics, Department of English, University of Birmingham, 2001, pp. 49–65.

Jakopin, P. 2001 b. Beseda : a Slovenian text corpus. In: FRASER, Michael (ed.), WILLIAMSON, Nigel (ed.), DEEGAN, Marilyn (ur.). *Digital Evidence : selected papers from DRH2000, Digital Resources for the Humanities Conference*, University of Sheffield, September 2000, (Office for Humanities Communication publication, 14). London: Office for Humanities Communication, 2001, pp. 229–41.

Sheehan, M. J. 2000. *Word Parts Dictionary: Standard and Reverse Listings of Prefixes, Suffixes, and Combining Forms*. Mc Farland & Co., Jefferson, NC, 227 pp.

Toporišič, J. 2000. *Slovenska slovnica*, *4., prenovljena in razširjena izd.* Obzorja, Maribor, 923 pp.

Vidovič-Muha, A. 1988. *Slovensko skladenjsko besedotvorje ob primerih zloženk*. Znanstveni inštitut Filozofske fakultete: Partizanska knjiga, Ljubljana:, 223 pp.

Žele, A. 2004. Stopnje terminologizacije v leksiki (na primerih glagolov) = The level of terminologization in the lexicon (in examples of verbs). V: HUMAR, Marjeta (ur.). *Terminologija v času globalizacije : zbornik prispevkov s simpozija Terminologija v času globalizacije, Ljubljana, 5.-6. junij 2003 : collected papers from the Scientific Conference Terminology at the Time of Globalization, Ljubljana, 5th-6th June 2003*. Ljubljana: Znanstvenoraziskovalni center SAZU, Založba ZRC: = Scientific Research Centre SASA, ZRC Publishing, pp. 77–91.

Žele, A. 2005. Novejša leksika z vidika aktualizacije pomenov in tvorbenih usmeritev. In: JESENŠEK, Marko (ur.). *Knjižno in narečno besedoslovje slovenskega jezika*, (Zbirka Zora, 32). Maribor: Slavistično društvo, pp. 240–48.