Query-driven Dictionary Enhancement

Primož Jakopin, Birte Lönneker
Corpus Laboratory
F.R. Institute of Slovenian language
ZRC SAZU, Novi Trg 2
1000 Ljubljana, Slovenia
primoz.jakopin@uni-lj.si, birte.loenneker@uni-hamburg.de

Abstract
With the log files of online dictionaries, in which all submitted user queries are stored, dictionary authors have for the first time in the history of dictionary building direct access to the users' requests. In this article, we show 1) how to use the log file to evaluate the current contents of an online dictionary and 2) how to choose the most promising corpus type for enlarging it according to the users' needs. We use the example of a German-Slovenian online dictionary for this work. As a result of the first evaluation, we detect that the dictionary does not fulfil the users' needs in coverage of colloquial and vulgar language, as well as in words and expressions used in everyday life. The result of the second evaluation confirms the importance of this part of the vocabulary. In an overall comparison of queries and corpora, a fiction corpus, which by its nature contains also colloquial language, yields a better result than a newspaper corpus and a non-fiction corpus.

1 Introduction

Print dictionaries allow only a limited evaluation against users' needs. Users of print dictionaries can send new words and suggestions to the editors, but not every user will do this every time he thinks a word might be missing, or an explanation unclear. The same limitation applies to CD-ROM editions of dictionaries. With the log files of online dictionaries, however, dictionary authors have for the first time access to the users' requests. In this article, we present the example of a German-Slovenian online dictionary (Section 2) and its log file (Section 3). Section 4 then shows how to use the log file to evaluate the current contents of the dictionary, and Section 5 presents a method of choosing the most promising corpus type for enlarging the dictionary according to the users' needs. Section 6 concludes the paper and discusses future work.

2 Online SLO-DE-SLO as an example of an online dictionary

Online SLO-DE-SLO is a bidirectional online dictionary for the language pair German-Slovenian. It started out as a learners' dictionary for German-speaking learners of Slovenian, containing the entire vocabulary of a Slovenian textbook for foreigners. This first collection was then completed by the 500 most frequent uncovered headwords occurring in an extensive newspaper corpus (Lönneker & Jakopin 2003). The version used for the subsequent evaluation contains 5,901 entries, comprising headwords, selected word forms and phrases; 729 entries of the dictionary correspond to the newspaper-based update. We can thus consider the evaluated version of Online SLO-DE-SLO as representing a slightly enlarged collection of basic Slovenian words and expressions with German counterparts.

Web interfaces to the dictionary exist in German and in Slovenian[1]; some additional information can be found in English. The monthly server statistics show that – based on the language of the user interface – the dictionary is about two to three times more popular among German-speaking users than among Slovenians, which reflects its initial purpose of making the Slovenian language more accessible to the German language community.[2] Users specify the translation direction (Slovenian into German or German into Slovenian) when entering their query. The matching of the query against the dictionary data can be performed in three different ways: exact match, string match at the beginning of a dictionary word or phrase, or string match anywhere inside the dictionary entry. All matches are case insensitive. For the convenience of the users, special characters can be substituted when entering a query: German umlauts by ae, oe, ue, German sharp s by ss, Slovenian diacritics (č, š, ž) by the corresponding basic ascii characters c, s, z. Special characters are most commonly replaced this way in these two languages.

3 Preprocessing of the log file

The log file available for the evaluation contains all 131,674 user queries to Online SLO-DE-SLO and its predecessors from 6 January 2002 to 10 October 2003. Out of these queries, 88,879 were performed using the exact string match option. Only these are retained for evaluation, as the other options support the submission of random parts of words, which are difficult to interpret.

The preprocessing involves downcasing of the selected queries (because of the case insensitive matching), splitting of the log file into Slovenian and German queries according to the direction the users chose for translation, and disambiguation of entries containing letters used for substituting special characters. The disambiguation is performed semi-automatically. For all queries containing one or more ambiguous characters or character sequences, we create possible parallel spellings; e.g., in the case of *fuer*, the parallel form *für* is created. Whether the ambiguous character or character sequence has to be replaced (in our example, *ue* by *ü*), is then determined by matching the parallel forms against the open-source ispell/aspell word list for German[3] or by comparing the frequencies of submitted queries in the parallel spelling systems (for Slovenian). Dubious cases are checked manually. A few Slovenian queries (e.g. *vas* for *vas* 'village' and *vaš* 'your') remain ambiguous and are retained in their "basic" spelling (e.g. *vas*). The last correction concerns the language of the submissions. Using the dictionary itself, already at this stage 378 queries are detected which are to be moved from the Slovenian to the German side, and 593 queries are moved from German to Slovenian. The resulting files used for evaluation hold 37,534 queries for Slovenian (into German) and 51,327 queries for German (into Slovenian).

4 Evaluation of the current dictionary

The dictionary file used for evaluation contains 5,901 entries (cf. Section 2). Due to polysemy, the number of distinct entries in each language is smaller. The German side contains 5,315 different entries (5,289 after downcasing), the Slovenian side 5,103.

For evaluation, the preprocessed queries are matched against the downcased dictionary entries in a straightforward way.[4] On the Slovenian side, 14,392 queries or 38,34% of the queries are successfully matched to a dictionary entry; 1,971 of them are *distinct*

queries/entries. Among the 23,142 Slovenian queries unknown to the dictionary, we find 13,527 distinct ones. On the German side, 20,907 queries or 40,73% are successfully matched to a dictionary entry; 1,840 of them are distinct. 30,420 German queries (15,779 distinct ones) are unknown to the dictionary.

Tables 1 and 2 show the Top 20 unmatched queries in Slovenian and German, respectively. We can see from the English translations that the dictionary lacks expressions and words used in social relations and everyday life, vulgar words, and terms for animals and plants. The list also contains some inflected forms (e.g. *rada*, *draga*, *gute*). It seems that the dictionary is often used in order to establish and maintain social contacts and for writing letters and messages.

#	Query	English translation
66	pozdrav	regard
65	ponudba	offer
42	potrdilo	confirmation; certificate
32	krava	cow
32	plačilo	payment
31	poljub	kiss
27	hrast	oak
27	pogrešati	to miss
26	rada	fond of (female form)
25	učiti	to teach
24	prodaja	sale
22	kokoš	hen
22	kraj	place
22	ljubim	I love
22	naročilo	order
22	zavarovanje	insurance
21	draga	dear (female form)
21	grozdje	grapes
20	davek	tax
20	dobrodošli	welcome (plural form)

Table 1: Top 20 unknown queries (Slovenian)

Obviously, the easiest way to improve dictionary performance would be to enter the most frequently submitted missing words (cf. Burke 1998:18). However, while this procedure might be successful for a dictionary of technical language such as the computer dictionary Burke refers to, a bilingual dictionary of general language should give users a broader background of the usage of words. For example, word-to-word translations of the

standard German expressions *bis bald* and *du fehlst mir* exist neither for Slovenian nor for English (meaning equivalences would be 'se vidiva/se vidimo' and 'pogrešam te' in Slovenian or 'see you soon'; 'I miss you' in English). A general dictionary should therefore not only provide translations of single words, but also illustrations of their use in different contexts, and the respective equivalents in the target language. The necessary contexts, from which phrasal usages and idioms can also be identified, could be found in a suitable corpus.

#	Query	English translation
145	kuß	kiss
130	willkommen	welcome
114	gruß	regard
112	grüße	regards
101	guten morgen	good morning
75	schatz	treasure
73	ficken	to fuck
72	guten abend	good evening
69	gute	good (female form)
67	sex	sex
67	vermissen	to miss
64	gute nacht	good night
57	arsch	ass
55	liebling	darling
54	vielen dank	thanks a lot
46	habe	have (1.sg.)
44	lieben	to love
43	bahnhof	train station
43	glückwunsch	congratulation
43	hase	bunny

Table 2: Top 20 unknown queries (German)

5 Choosing a corpus for dictionary enlargement

The utility of corpora for lexicographical work in general is probably indisputable since the successful completion of seminal monolingual dictionary projects like COBUILD for English (Sinclair 1987) and TLF for French (Gorcy 1992). A more elaborate way of improving Online SLO-DE-SLO than that of adding single missing words would thus consist in choosing a corpus that best reflects the *structure of the entire vocabulary* entered by the users, and to cover the words, expressions and idioms in that corpus. In this section, we will present a way of choosing such a corpus.

Considering the background of the dictionary, which is to provide German-speaking people a better access to the Slovenian language and culture, we decided to perform this part of the evaluation using Slovenian corpora only. In contrast to Gorjanc & Krek (2001), we are not interested in the most frequent words of a balanced corpus of Slovenian, but in choosing a specific Slovenian (sub)corpus which covers the needs of the dictionary users best, and from which words and expressions should be extracted for inclusion in the dictionary. For this task, we will compare all lemmas in the user queries with relative frequencies of lemmas

in three Slovenian corpora of different text types (fiction, newspaper and non-fiction), which are actually subcorpora of the 100 million word corpus *Nova Beseda* (Jakopin 2001).

The lemmatization of the 15,498 distinct Slovenian queries is performed in the following steps:

1. The log file prepared according to the description in Section 3 is lemmatized using the lemmatizer of the Institute for Slovenian Language[5]. 9,213 queries can be lemmatized.

2. Unlemmatized queries that contain at least one blank (2,171 distinct queries) are split into 6,026 single words (3,304 distinct word forms), which are then lemmatized as well. 2,550 forms are lemmatized, while 754 forms cannot be lemmatized in this step.

3. From the list of the 4,868 word forms that remain unlemmatized at this stage, all *hapax legomena* are discarded, because due to spelling errors or to non-Slovenian words, the majority of them would be difficult to treat (cf. examples in Table 3).

Query
bettonug
bezen
bicikl
bijal
bijc
bikoviny
bilti

Table 3: Examples of *hapax legomena*

4. The remaining unlemmatized 972 queries are manually checked and categorized into different groups. In 436 cases, an erroneous spelling of an existing Slovenian word can be detected, while 72 queries are correct Slovenian words unknown to the lemmatizer, and 50 are colloquial Slovenian words. Other major identified groups of unlemmatized queries are those of German words, Slovenian proper names, and Croatian, Czech, and Polish words, in descending order of frequency. These items were not further processed.

5. After a correction of the mistakes in the manually checked list and an adaptation of the lemmatizer with new or colloquial Slovenian words, all these queries can be lemmatized. The total list of lemmatized Slovenian queries from all previously mentioned steps contains 10,679 distinct lemmas corresponding to 37,632 word forms.

6. From this lemma list, only content words (nouns, verbs, adjectives, and adverbs; cf. open word classes in Greenbaum (1996)) and interjections are retained for evaluation. For identical lemmas to which more than one part of speech (POS) was assigned by the lemmatizer, only one POS is selected. This is done using some heuristics; for example, the POS-combination 'ADV,ADJ' is represented by 'ADJ' only, because the adverb is derived from the corresponding adjective. 40 lemmatized queries cannot be disambiguated and are discarded.

7. The final list of Slovenian lemmas for evaluation contains 7,246 lemmas with their POS-tags.

Our next task consists in finding relative frequencies of lemmas in the three different candidate corpora: a fiction corpus (Slovenian fiction from early 20th to 21st century; ca.

5,677,000 words), a newspaper corpus (Slovenian daily DELO; ca. 88,426,000 words), and a non-fiction corpus (essays, letters, scientific monographs, technical and health magazines; ca. 6,273,000 words). The corpora are lemmatized using the updated version of the lemmatizer with an integrated guesser for unknown words. A POS-disambiguation is performed for identical lemmas with different POS according to the heuristics already used for query treatment. In case of other ambiguities, the corpus frequency is assigned to both lemmas. The *numbers of distinct lemmas* for each corpus are as follows:

1. Fiction 71,298
2. Newspaper 352,898
3. Non-fiction 94,241

Instead of total occurrences, relative frequencies (per million words after lemmatization) are assigned to each lemma because of the different size of the corpora.[6] Treating each corpus separately, we then multiply, for each lemma, the number of logged dictionary queries with its relative frequency in the corpus, which results in a certain "weight" for each lemma representing its importance both in the corpus and in the user queries. Table 4 shows the first seven lines of the fiction corpus evaluation, in alphabetical order; Table 5 displays the top 20 weighted lemmas of the same corpus. Analogous tables for the other corpora can be found in the appendix. The top 20 weighted lemma tables show a great deal of overlap in lemmas; those lemmas occurring in the top 20 list of only one of the corpora appear in bold font in Table 5 and in the respective Tables 9 and 10 in the appendix.

Lemma	English	Frequency p.m.	# Queries	Weight
absoluten:P	absolute	7.64	1	7.64
absolvent:S	graduate	0.28	2	0.56
adaptacija:S	adaptation	0.14	2	0.28
adrenalin:S	adrenalin	1.13	1	1.13
afera:S	affair	1.56	3	4.68
afna:S	ape; at-sign	0.14	3	0.42
agencija:S	agency	6.65	13	86.45

Table 4: First seven lines of fiction corpus evaluation (P = adjective; S = noun)

Lemma	English	Frequency p.m.	# Queries	Weight
biti:G	to be	7695.06	387	2977988.22
imeti:G	to have	12822.89	149	1910610.61
dati:G	to give	13658.78	29	396104.62
iti:G	to go	2122.76	154	326905.04
dober:P	good	1292.82	234	302519.88
dan:S	day	1284.33	194	249160.02
hiša:S	house	828.81	300	248643.00
lep:P	beautiful	1085.93	218	236732.74
miza:S	table	664.52	216	143536.32
priti:G	to come	1692.01	81	137052.81
lahek:P	light	1347.72	84	113208.48
vedeti:G	to know	2060.92	37	76254.04
videti:G	to see	1934.70	35	67714.50
misliti:G	to think	1506.21	44	66273.24

reči:G	to say	2727.14	24	65451.36
gledati:G	to look	1334.99	48	64079.52
velik:P	big	1725.55	36	62119.80
rad:P	fond of	704.57	87	61297.59
leto:S	year	862.77	64	55217.28
delati:G	to work; to do	605.51	88	53284.88

Table 5: Top 20 weighted lemmas from fiction corpus (P = adjective; S = noun; G = verb)

The last step in corpus evaluation consists in adding up the resulting weights of all lemmas for each corpus. The results are as follows:

1. Fiction 10,262,558.41
2. Newspaper 9,694,125.71
3. Non-fiction 9,369,494.16

Even though the results for all three corpora are of the same order of magnitude, the difference between the fiction and the newspaper corpus is noticeable enough to say that a fiction corpus fits the user queries better than other corpora.

6 Conclusion and discussion

The advantage of an online dictionary is that the queries can be logged and used in order to evaluate and enhance the dictionary. In this article, we showed how the queries of a bilingual online dictionary are mapped onto the dictionary contents and onto three corpora representing different text types. As a result of the first evaluation, we detect that the dictionary does not fulfil the users' needs in coverage of colloquial and vulgar language, as well as in words and expressions used in everyday life. The result of the corpus-based experiment confirms the importance of this part of the vocabulary. In an overall comparison of queries and corpora, the fiction corpus, which by its nature contains also colloquial language, yields a better result than the newspaper corpus and the non-fiction corpus. According to these findings, the evaluated Online SLO-DE-SLO dictionary should thus be enhanced with words and expressions from a fiction corpus, if it wants to serve its users best. This result contradicts the current common practice for online learners' dictionaries (Erjavec, Hmeljak Sangawa & Srdanović 2003; Lönneker & Jakopin 2003), which is to focus on newspaper coverage.

We believe that our idea of selecting corpora for dictionary enlargement on the grounds of user requests can be of use for a large variety of online dictionaries. We would therefore like to give some directions of how our procedure of corpus evaluation could further be modified and experimented on. For example, future work could determine the eventual usefulness of function words in the analysis method. In the present study, we excluded them from the evaluation because they do not directly convey "information". Lexical density (as the relative proportion of content to function words in a text) is however sometimes considered to be an indicator of text type: Stubbs (1996:71–73) assumes that a low lexical density shows a tendency towards informal speech allowing feedback, while a high lexical density suggests a variety of formal text (spoken or written). Function words have the additional advantage of showing relatively accurate probability estimates already in a corpus

of 5 million words, in an experiment that Curran & Osborne (2002) carried out on a 1.145 billion words corpus of English newspaper text.

On the other hand, small corpora and subcorpora are not a problem for our method, but might even be desired in online dictionary creation. Taking advantage of the "burstiness" of words (Curran & Osborne 2002:129), we could say that texts are good sources for online dictionary enlargement if frequent query words "burst" in them. A natural continuation of the procedure could thus be to further subdivide the most suitable corpus (in our case, the fiction corpus) into different parts (for example, works by different authors) and to repeat the experiment on them. Some of the most successful shorter texts could then be selected, and their entire vocabulary be covered by the online dictionary. Optimally, with the consent of the copyright owners, the texts would be made available on the dictionary homepage. This would give language learners the opportunity to study entire coherent texts containing useful expressions, with full support by the dictionary.

Acknowledgements

The work was partly supported by the project "Fostering a Knowledge-based Society in Slovenia" funded by the United Nations Development Programme and the Republic of Slovenia, implemented by the Slovenian Science Foundation. We also thank Katarina Rozman who, in 2002–2004, verified the dictionary entries corresponding to the textbook. This dictionary proofreading was funded by the Laurence Urdang EURALEX awarded to Birte Lönneker in 2001.

Endnotes

1 URLs: http://www.rrz.uni-hamburg.de/slowenisch (German interface); http://www.rrz.uni-hamburg.de/slowenisch/index_si.htm (Slovenian interface) [last accessed on 2 March 2004].

2 In this regard, the dictionary differs from a similar one described by Erjavec, Hmeljak Sangawa & Srdanović (2003), which is meant for Slovenian learners of Japanese.

3 The file `all.words` from `igerman98-20021114` contains more than 146,000 entries (URL: http://j3e.de/ispell/igerman98/dict/ [27 October 2003]).

4 Downcasing is necessary, as the log file has been normalized in the same way (cf. Section 3).

5 URL: http://bos.zrc-sazu.si/dol_lem.html [27 October 2003].

6 The three corpora are of different size (text length); the largest corpus provides also the biggest number of lemmas (types). We would like to point out that the logarithmic type/token ratio (Herdan 1960:26–33) – used as a quantitative expression for the tendency of the number of types to grow with corpus size – is comparable for the three corpora (cf. last column in Table 6).

Corpus	# Words	# Lemmas	Type/Token ratio	γ (log V/log N)
Fiction	5,677,000	71,298	0.0126	0.718535967
Newspaper	88,426,000	352,898	0.0039	0.698117834
Non-fiction	6,273,000	94,241	0.015	0.731777561

Table 6: Type-token and bilogarithmic type-token ratio for the three evaluated corpora.

References

Burke, S. M. 1998. *The Design of Online Lexicons.* Master's thesis: Northwestern University, Evanston, IL.

Curran, J. R. and Osborne, M. 2002. 'A very very large corpus doesn't always yield reliable estimates' in *Proceedings 6th Conference on Natural Language Learning (CoNLL-2002), Taipei, Taiwan, 31 August – 1 September, 2002.* 126–131.

Erjavec, T., Hmeljak Sangawa, K. and Srdanović, I. 2003. 'An XML TEI Encoding of a Japanese-Slovene Learners' Dictionary' in *Proceedings 6th International Multiconference Information Society 2003 (IS-03), Ljubljana, Slovenia, 13–17 October 2003.* B: 20–26.

Gorcy, G. 1992. 'Le Trésor de la langue française (TLF) trente ans après; bilan et perspectives' in *Études de linguistique appliquée 85/86.* 75–88.

Gorjanc, V. and Krek, S. 2001. 'A Corpus-based Dictionary Database as the Source for Compiling Slovene-X dictionaries' in *Proceedings 6th Conference on Computational Lexicography and Corpus Research "Computational Lexicography and New EU Languages" (COMPLEX 2001), Birmingham, 28 June – 1 July, 2001.* 41–47.

Greenbaum, S. 1996. *The Oxford English Grammar.* Oxford: Oxford University Press.

Herdan, G. 1960. *Type-token mathematics: A textbook of mathematical linguistics.* The Hague: Mouton.

Jakopin, P. 2001. '*Beseda*: a Slovenian Text Corpus' in M. Fraser et al. (eds.), *Digital Evidence: Selected Papers from DRH2000, Digital Resources for the Humanities Conference, University of Sheffield, September 2000.* London: Office for Humanities Communication. 229–241.

Lönneker, B. and Jakopin, P. 2003. 'Contents and Evaluation of the First German-Slovenian Online Dictionary' in *10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003), Budapest, Hungary, Conference Companion.* 119–122.

Sinclair, J. M. (ed.) 1987. *Looking up. An account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary.* London/Glasgow: Collins ELT.

Stubbs, M. 1996. *Text and Corpus analysis. Computer-assisted Studies of Language and Culture.* Oxford: Blackwell.

Appendix

A: First seven lines of corpus evaluation (alphabetically)

Lemma	English	Frequency p.m.	# Queries	Weight
absoluten:P	absolute	7.64	1	7.64
absolvent:S	graduate	0.28	2	0.56
adaptacija:S	adaptation	0.14	2	0.28
adrenalin:S	adrenalin	1.13	1	1.13
afera:S	affair	1.56	3	4.68
afna:S	ape; at-sign	0.14	3	0.42
agencija:S	agency	6.65	13	86.45

Table 7: Fiction corpus (P = adjective; S = noun)

Lemma	English	Frequency p.m.	# Queries	Weight
absoluten:P	absolute	46.18	1	46.18
absolvent:S	graduate	1.82	2	3.64

adaptacija:S	adaptation	4.38	2	8.76
adapter:S	adapter	0.27	1	0.27
adaptirati:G	to adapt	0.75	1	0.75
adjektiv:S	adjective	0.01	1	0.01
adrenalin:S	adrenalin	2.95	1	2.95

Table 8: Newspaper corpus (P = adjective; S = noun; G = verb)

Lemma	English	Frequency p.m.	# Queries	Weight
absoluten:P	absolute	61.69	1	61.69
absolvent:S	graduate	0.89	2	1.78
adaptacija:S	adaptation	1.04	2	2.08
adapter:S	adapter	2.53	1	2.53
adaptirati:G	to adapt	0.45	1	0.45
adjektiv:S	adjective	0.30	1	0.30
adrenalin:S	adrenalin	2.68	1	2.68

Table 9: Non-fiction corpus (P = adjective; S = noun; G = verb)

B: Top 20 weighted lemmas of corpus evaluation

Lemma	English	Frequency p.m.	# Queries	Weight
biti:G	to be	7695.06	387	2977988.22
imeti:G	to have	12822.89	149	1910610.61
dati:G	to give	13658.78	29	396104.62
iti:G	to go	2122.76	154	326905.04
dober:P	good	1292.82	234	302519.88
dan:S	day	1284.33	194	249160.02
hiša:S	house	828.81	300	248643.00
lep:P	beautiful	1085.93	218	236732.74
miza:S	table	664.52	216	143536.32
priti:G	to come	1692.01	81	137052.81
lahek:P	light	1347.72	84	113208.48
vedeti:G	to know	2060.92	37	76254.04
videti:G	to see	1934.70	35	67714.50
misliti:G	to think	1506.21	44	66273.24
reči:G	to say	2727.14	24	65451.36
gledati:G	to look	1334.99	48	64079.52
velik:P	big	1725.55	36	62119.80
rad:P	fond of	704.57	87	61297.59
leto:S	year	862.77	64	55217.28
delati:G	to work; to do	605.51	88	53284.88

Table 10: Fiction corpus (P = adjective; S = noun; G = verb)

Lemma	English	Frequency p.m.	# Queries	Weight
imeti:G	to have	17457.81	149	2601213.69
biti:G	to be	4564.92	387	1766624.04
dober:P	good	1573.44	234	368184.96
dati:G	to give	11812.20	29	342553.80
dan:S	day	1406.23	194	272808.62
leto:S	year	3985.21	64	255053.44
lahek:P	light	2240.81	84	188228.04
iti:G	to go	1077.23	154	165893.42
delo:S	work	2083.23	60	124993.80
hiša:S	house	330.57	300	99171.00
velik:P	big	2729.40	36	98258.40
podjetje:S	company	887.86	82	72804.52
delati:G	to work; to do	827.14	88	72788.32
mesto:S	town; place	1204.45	56	67449.20
lep:P	beautiful	282.88	218	61667.84
nov:P	new	2189.42	27	59114.34
priti:G	to come	716.19	81	58011.39
morati:G	to have to	1641.29	34	55803.86
slovenski:P	Slovenian	1954.48	24	46907.52
del:S	part	1618.73	28	45324.44

Table 11: Newspaper corpus (P = adjective; S = noun; G = verb)

Lemma	English	Frequency p.m.	# Queries	Weight
imeti:G	to have	15939.66	149	2375009.34
biti:G	to be	3526.60	387	1364794.20
dober:P	good	1752.50	234	410085.00
lahek:P	light	4776.29	84	401208.36
dati:G	to give	11110.61	29	322207.69
delo:S	work	2593.13	60	155587.80
leto:S	year	2428.01	64	155392.64
iti:G	to go	997.55	154	153622.70
lep:P	beautiful	577.01	218	125788.18
dan:S	day	641.99	194	124546.06
podjetje:S	company	1482.17	82	121537.94
velik:P	big	3366.55	36	121195.80
delati:G	to work; to do	899.64	88	79168.32
nov:P	new	2711.00	27	73197.00
računalnik:S	computer	2294.49	31	71129.19
slika:S	picture	1308.41	48	62803.68
priti:G	to come	712.32	81	57697.92
del:S	part	2055.75	28	57561.00
podatek:S	data item	1665.91	34	56640.94
morati:G	to have to	1608.99	34	54705.66

Table 12: Non-fiction corpus (P = adjective; S = noun; G = verb)