Slovene Word Sketches

Simon Krek,[*] Adam Kilgarriff,[**]

[*] Faculty of Arts
University of Ljubljana
Ljubljana, Slovenia
simon.krek@guest.arnes.si

[**] Lexical Computing Ltd
Brighton, United Kingdom
adam@lexmasterclass.com

Abstract

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. Today, 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 FidaPLUS corpus, a morpho-syntactically tagged corpus of Slovene was loaded into the Sketch Engine software. We shall demonstrate the Slovene word sketches, and show how they can be used in lexicography and for other linguistic purposes. The results show that word sketches could significantly facilitate lexicographic work in Slovene as they have for English.

Povzetek

Besedne skice *(Word sketches)* so avtomatski na korpusu temeljeci sežetki slovnicnega in kolokacijskega vedenja neke besede. Prvic so bile uporabljene pri sestavljanju enojezicnega angleškega slovarja založbe Macmillan (Rundell 2002). Takrat so obstajale le za angleški jezik. Zdaj je na voljo programski modul Sketch Engine, korpusno orodje, ki na vhodu sprejme korpus kateregakoli jezika ter njegove slovnicne vzorce, iz njih pa ustvari besedne skice za besede tega jezika. Hkrati avtomatsko generira tezaver in "razlikovalne skice", ki izpostavljajo podobnosti in razlike med bližnjimi sopomenkami. V programski modul Sketch Engine smo naložili korpus FidaPLUS, oblikoslovno-skladenjsko oznaceni korpus slovenšcine. Prikazali bomo slovenske besedne skice in pokazali, kako jih je mogoce uprabiti za leksikografske in druge jezikoslovne namene. Rezultati kažejo, da besedne skice znatno olajšajo delo leksikografom slovenskega jezika, tako kot se je izkazalo pri angleščini.

.

1. Introduction

Word sketches are one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour. Their value for lexicographic work in English and other languages, as well as the background of the use of corpora in lexicography, have been described elsewhere (Kilgarriff and Tugwell 2001, Kilgarriff and Rundell 2002, Kilgarriff et al. 2004).

First, we shall introduce corpus query systems and the basic idea of word sketches. Next, we shall concentrate on the application of word sketches to the Slovene language in the Sketch Engine software.

The FidaPLUS corpus of Slovene will also be briefly described, with special attention to the tagging problems which could affect its use within the Sketch Engine.

2. Word sketches

2.1. Corpus query systems

Different corpus query systems have been used to check the corpus evidence since the rise of the first electronic corpora. Ever since the COBUILD project, lexicographers have been using KWIC concordances as their primary tool for finding out how a word behaves. Later, with the growth of corpora, lexical statistics had to be applied to manage the abundant data and highlight the most salient combinations and collocations. Today, state-of-the-art CQSs allow the lexicographer great flexibility in searching for phrases, collocates, grammatical patterns, sorting concordances according to a wide range of criteria, identifying 'subcorpora' for searching in only spoken text, or only fiction. Available systems include WordSmith, MonoConc, and the Stuttgart Workbench among others.

Specifically for the two large Slovene corpora, there are also two different on-line concordancers available: ASP32 for the FidaPLUS corpus[1] and NEVA for Nova beseda, with a more detailed description available in Krek (2003).[2]

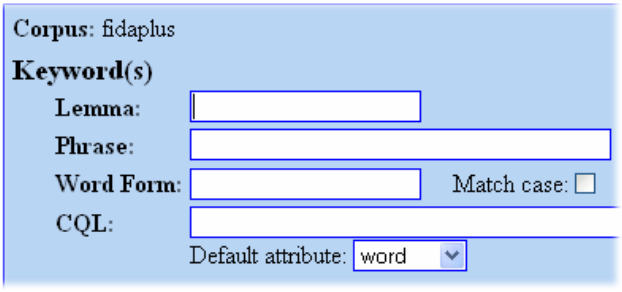
2.2. Sketch Engine

2.2.1. Description

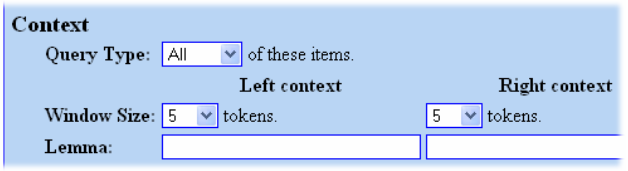
The Sketch Engine is a corpus query system which allows the user to use the familiar CQS functions:

[1] http://www.fidaplus.net
[2] http://bos.zrc-sazu.si/s_beseda.html

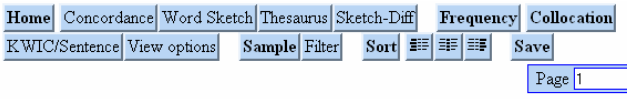
– concordances with lemma, phrase, word form and CQL search,



together with the context control filter



and the usual viewing and sorting options:



F0000012.35.10 polkrožna platoja . Danes tam stoji počitniška luša , ob njej pa zidanic
F0000012.104.9 vgnezdila nemška posadka . Zdaj tam stoji nova luša . Grajska kapela M
F0000012.214.1 zgodovinski listini omenjena Karolova " luša " (haws Sagradez
F0000012.224.7 , ohranjena je še nekdanja oskrbnikova luša , nekdanja grajska
F0000012.295.5 Stančič , zdaj pa na njenem mestu stoji nova luša . Južno od tod so p
F0000012.614.8 Ob grajskem jedru je bila zgrajena nova luša . Dornberk je bil ol
F0000012.782.8 podrl stavbo . Na njegovem mestu stoji nova luša . Galetov
F0000012.1001.16 Buka . Danes na njenem mestu stoji nova luša . Gospodarsko pos

However, the features of the Sketch Engine which are of special interest in this article are not part of standard concordancing programs. These features include Word Sketch, Sketch Difference and Thesaurus which will be described later. All these features are fully integrated with standard concordancing.

2.2.2. Word Sketch

To identify a word's grammatical and collocational behaviour, the Sketch Engine needs to know how to find words connected by a grammatical relation. It allows two possibilities.

In the first, the input corpus has been parsed and the information about which word-instances stand in which grammatical relations with which other word-instances is embedded in the corpus. Currently, dependency-based syntactically annotated corpora are supported. Phrase-structured trees need heads of phrases to be marked.

In the second, the input corpus is loaded into the sketch engine POS-tagged but not parsed, and the sketch engine supports the process of identifying grammatical relation instances. Each grammatical relation will be defined, using the Sketch Engine to test and develop it. When the developer is happy with the definition of each grammatical relation, they save the definitions in a "gramrel" file. The Sketch Engine then compiles this file and finds all instances of all grammatical relations in the corpus. It puts them in a gramrels database and users than have access to word sketches.

2.2.3. Lemmatization & POS-tagging

The Sketch Engine does not support the process of lemmatization; various tools are available for linguists to develop lemmatizers, and they are available for a number of languages. If no lemmatizer is available, it is possible to apply the Sketch Engine to word forms, which, while not optimal, will still be a useful lexicographic tool.

Similarly for part of speech (POS) tagging, also known as POS-disambiguation. This is the task of deciding the correct word class for each word in the corpus – of determining whether an occurrence of "brez" in Slovene is an occurrence of a noun "breza" in plural, genitive case, or a preposition. A tagger presupposes a linguistic analysis of the language which has given rise to a set of the syntactic categories of the language, or tagset. Tagsets and taggers exist for a number of languages, and there are assorted well-tried methods for developing taggers. The Sketch Engine assumes tagged input.

As the FidaPLUS corpus is both lemmatized and POS-tagged but not syntactically annotated, Slovene word sketches are based on a lemmatized and POS-tagged corpus, with grammatical relations defined on the basis of POS-tag information.

2.3. Grammatical relations

Grammatical relations are defined as regular expression over POS-tags. For example, **if** we wish to include the grammatical relation between a noun and its adjectives in modifying position, we define the head of the noun phrase, a noun ("S" in the FidaPLUS tagset) and one or more preceding adjectives ("P") with the possibility of allowing the intervening comma and the particles "se" and "si":

=a_modifier/modifies
2: [tag="P.*"] [tag="P.*" | word="," | word="se" | word="si"] {0,5} 1: [tag="S.*"]

The first line, following the =, gives two names for the grammatical relation. The first, before the slash, is the name when the arguments are in the one order, and the other is when the arguments are in the other.

The 1: and 2: mark the words to be extracted as the first and second arguments. |, ., (), and * are standard regular expression metacharacters. {0,5} indicates that the preceding term occurs between zero and five times.

3. Slovene Word Sketches

3.1. Slovene Corpus

3.1.1. FIDA corpus

The FIDA corpus is the precursor of the FidaPLUS corpus which was used in the Sketch Engine software. It was compiled in a joint project involving four partners, two from the academic/research sphere: (the Faculty of Arts, University of Ljubljana, the Jožef Stefan Institute) and two commercial ones (DZS publishing house and Amebis software company). Corpus compilation started in 1997 and was concluded in 2000. The corpus was just over 100 million words and was a balanced corpus of texts in the Slovene language mainly from the 1990s.

The corpus was lemmatized and POS-tagged but the process was limited to the lexicon of word forms available

at Amebis at the time. The disambiguation of multiple possible morphosyntactic descriptions, (MSDs) for ambiguous wordforms such as *brez* was not performed, a considerable drawback when using the corpus for automatic linguistic analysis.

3.1.2. FidaPLUS corpus

The problems of lemmatization and POS-tagging, together with the size, balance and up-to-dateness were addressed in the subsequent project, "Language Resources for Slovene", funded by the Slovene Ministry of Higher Education, Science and Technology and co-funded by DZS and Amebis. Project partners included the Faculty of Arts (University of Ljubljana) as the leading partner, the Faculty of Social Sciences (University of Ljubljana) and the Jožef Stefan Institute. Its aim was a three hundred million word corpus with complete lemmatization and POS-tagging.

The FidaPLUS corpus used for testing in the Sketch Engine is the preliminary result of the project. In terms of size it is similar to the FIDA corpus, but the lemmatization and POS-tagging have been improved. Lemmatization is both lexicon-based and statistical, aiming at lemmatization of all items in the corpus. POS-disambiguation uses the tools developed by Amebis.

3.2. Slovene grammatical relations

The Slovene "gramrel" file was based on the Czech example (Kilgarriff et al. 2004), since Czech, like Slovene but unlike English, is a relatively free word order language.

The grammatical relations in the Slovene gramrel file include three types: **symmetric**, between two items with equal status, **dual**, between two items with dependent relations and **trinary**, between three dependent items.

coord	7334	0.5
prostor	837	44.63
datum	144	37.89
trud	75	35.92
kraj	247	34.76
Vera	33	31.35
energija	155	30.85
denar	206	27.91
se	192	26.5
bit	35	25.96
zaslonka	13	25.71
potrpljenje	21	24.85
trimesečen	17	23.36
da	97	23.18
on	131	21.66
svoj	26	20.95
ne	63	19.83
napor	25	19.0
tudi	43	18.33
kraja	24	18.02
njegov	35	17.97

3.2.1. Symmetric Example

One example of the symmetric relation is various coordinate structures with conjunctions "and" or "or", as well as two-word coordinate structures such as "niti-niti", "ali-ali".

```
=coord
*SYMMETRIC
        1:[] [word = "in" | word = "ali"] 2:[]
        [word = "niti"] 1:[] [word = "niti"] 2:[]
        [word = "ali"] 1:[] [word = "ali"] 2:[]
        [word = "bodisi"] 1:[] [word = "bodisi"] 2:[]
        [word = "tako"] 1:[] [word = "kakor"] 2:[]
        [word = "tako"] 1:[] [word = "kot"] 2:[]
```

The result of this grammatical relation can be viewed as part of the word sketch. The result shows that in the FidaPLUS corpus, 7334 instances of this particular grammatical relation can be found for the lemma "cas". Lemmas are ranked according to the salience score (Kilgarriff and Tugwell 2001). The user can click on the number next to a lemma to see the relevant concordance.

We used four summetrical relations..

a_modifier	19068	1.4
sklonjen	157	59.54
obrit	66	47.71
kronan	35	39.24
dvignjen	62	36.12
zeljnat	23	35.75
odsekan	33	35.5
Hermanov	34	35.11
razgret	40	34.63
mrtvaški	34	31.12
trezen	41	30.65
bikov	22	29.33
ovnov	16	29.12
video	106	29.03
koničast	34	28.96
pobrit	13	27.94
zeljen	18	27.92
bister	35	27.26

is_obj4_of	2506	3.9
skloniti	98	54.21
beliti	83	49.81
dvigniti	218	49.15
razbijati	71	44.74
odsekati	60	43.2
pomoliti	45	42.62
nagniti	59	40.97
tiščati	46	40.92
stakniti	39	39.75
obrniti	90	34.92
odrezati	44	33.54
nasloniti	29	32.21
sploščiti	20	32.18
razbiti	36	31.43
pobešati	11	30.46
sklanjati	18	30.17
povešati	11	29.91

3.2.2. Dual Example

Dual relations are most common in the gramrel file. There are eleven of them, covering relations expressed by means of grammatical case in Slovene as well as modifying structures as shown before. The corresponding part of the word sketch for the lemma "glava" is shown on the left.

Relations covering grammatical cases are defined in the following fashion:

```
=is_obj4_of/has_obj4
*DUAL
        2:[tag="Gpp.*" & !(lemma = "biti" | lemma = "imeti" | lemma = "hoteti" | lemma = "morati" | lemma = "smeti") ] [tag!="[SGDVLMOZ].*" & tag!=""] {0,5} 1:[tag="S...t.*"]
        2:[tag="G.d.*" & !(lemma = "biti" | lemma = "imeti" | lemma = "hoteti" | lemma = "morati" | lemma = "smeti") ] [tag!="[SGDVLMOZ].*" & tag!=""] {0,5} 1:[tag="S...t.*"]
```

There are two variants of the particular relation: either a verb has an object in the oblique case or the noun is itself an object in the same case, in relation to a verb. The example on the left shows a list of verbs where the lemma "glava" is predominantly used in the oblique case within a window of five items from a verb. All the verbs from the beginning of the list indicate structures which are lexicographically relevant because of their either central or additional metaphorical meaning. Thus the concordances of the structure "skloniti glavo" show that besides the literal meaning "to bow one's head", there are many examples of the metaphorical extension "to give up" or "to concede defeat". The next one indicates the structure "beliti si glavo" which is thoroughly idiomatic: "to worry about, to agonize over". The same is true for "razbijati si glavo", "tiščati glave (skupaj)", "stakniti glave" etc.

3.2.3. Trinary Example

Trinary relations indicate the relations between three grammatical categories. In the Slovene gramrel file, they are mainly used to extract prepositional patterns where the grammatical case – in Slovene the instrumental and locative cases – is expressed by means of prepositional phrases.

prec_po	1090	9.0
rojiti	91	65.67
udariti	142	55.22
motati	49	51.25
popraskati	24	42.95
treščiti	38	39.83
poditi	32	38.99
tolči	34	37.54
lopniti	12	32.97
tepsti	20	32.27
praskati	13	28.88
bloditi	12	28.06
plesti	14	27.2
trepljati	9	26.25
poškodovati	23	25.79
čohati	6	25.61
pobriti	5	24.37
srati	8	23.96

*TRINARY
=prec_%s
2:[tag="S.*"] 3:[tag="D.*"] [tag="P.*" | word= "," | word = "se" | word = "si"] {0,5} 1:[tag="S.*"]
2:[tag="G.*"] 3:[tag="D.*"] [tag="P.*" | word= "," | word= "se" | word="si"] {0,5} 1:[tag="S.*"]

In the case shown on the left, the grammatical relation is established between the lemma "glava" preceded by the preposition "po", and the "glava" word sketch indicates salient combinations with verbs on the left. Again, together with the frequent but semantically transparent combinations there are numerous idiomatic expressions such as "rojiti/motati/poditi po glavi" and the more informal "srati po glavi".

3.3. Sketch Differences

The sketch differences feature in the Sketch Engine specifies, for two semantically related words, what behaviour they share and how they differ. Synonymous words tend to share some of the collocates but not all. The sketch differences show the patterns which are shared by both synonyms and presents the information also in a colour scheme for the user to grasp immediately if and where the lemmas are synonymous. For the Slovene language, this is particularly useful in cases where there are two competing synonyms, one etymologically foreign and the other of Slavic origin. The more normatively-minded usually argue for abolition of the foreign lemma and non-discriminatory use of the Slavic form. The example of "cona" and "obmocje" in the Appendix 1 shows the differences. In the FidaPLUS corpus, only "operativen" is distributed evenly between the two synonyms. A milder bias towards "obmocje" is indicated in the cases of "demilitariziran" and "turisticen" and a stronger one with "zaprt" and "obmejen". The opposite is true with more fixed "erogena cona", "obrtna cona", "industrijska cona" etc. and less fixed "carinska cona / carinsko obmocje", "tamponska cona / tamponsko obmocje", also "tamponski", "brezcarinski", "siv" etc.

3.4. Thesaurus

The similarity is based on 'shared triples'. "Cona", "obmocje" both occur as the second term in the triple <modifier, ?, "tamponska">, and this provides one small piece of evidence that the two words are close in meaning. By simply gathering together all such pieces of evidence (and weighting them according to salience, following the method developed by Lin (1998)), we identify the near neighbours for each. The Sketch Engine does this and the result for the lemma "kriza" can be seen in the Appendix 2.

As there is no thesaurus available for the Slovene language, it is not possible to compare it to the human assessment of the word's synonymic relations, but it is immediately clear that the software shows a number of relevant items such as "konflikt", "spor", "spopad" etc., indicating one semantic direction, "problem", "težava", "zaplet" etc., indicating another, and "stiska", "izguba" indicating a more intimate human sentiment.

One can explore each of the relations with the sketch differences feature.

4. Conclusion and further work

Testing of the 100-million FidaPLUS corpus in the Sketch Engine has shown it to be an exceptionally useful tool for exploring typical grammatical and lexical relations in the Slovene language. To be able to take full advantage of the software, it is important to have a corpus which is lemmatized and POS-tagged as accurately as possible, and that is one area where there is room for improvement. We would like to further explore Slovene grammatical relations and their implementation in the gramrel file, and also the possibility a Slovene dependency-parser.

However even in its present form the Sketch Engine is a valuable tool, particularly for lexicographic use.

5. References

Kilgarriff, A., Tugwell, D. (2001). WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. *Proc. ACL workshop on COLLOCATION: Computational Extraction, Analysis and Exploitation*. Toulouse. 32-28.

Kilgarriff, A., Rundell, M. (2002). Lexical profiling software and its lexicographic applications - a case study. *Proc EURALEX*. Copenhagen. 807-818.

Kilgarriff, A., Rychly, P., Smrž, P., Tugwell, D. (2004) The Sketch Engine. *Proc. Euralex*. Lorient, France. 105-116.

Lin, D, (1998). Automatic retrieval and clustering of similar words. *COLING-ACL*, Montreal. 768-774.

Krek, S. (2003). Jezikovni prirocniki in novi mediji. *Jezik in slovstvo*, letn. 48, št. 3-4, 29-46.

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

Appendix 1: Sketch difference – lemma_1 “cona”, letmma_2 “obmocje”

	a_modifier	4941	36173	2.6	1.9
green	erogen	75	13	58.6	21.4
green	obrten	288	9	58.3	5.6
green	ekonomski	411	19	54.5	5.2
light green	carinski	271	214	53.4	33.9
green	industrijski	284	76	52.2	19.4
green	prost	317	79	50.1	16.7
green	prostocarinski	61	21	47.3	21.8
red	obmejen	7	313	11.5	46.7
green	moder	139	20	41.4	8.2
green	okupacijski	35	10	38.6	14.6
light red	nekdanji	41	710	16.6	37.3
extra light green	tamponski	17	12	35.8	23.1
extra light red	demilitariziran	9	30	25.3	33.8
red	zaprt	5	162	7.6	32.4
extra light green	brezcarinski	26	17	30.3	16.5
light green	konvergenčen	12	6	29.7	15.0
white	operativen	34	47	28.5	21.8
light red	bivši	15	199	11.9	27.4
red	posamezen	9	338	5.1	26.8
light green	svoboden	42	43	25.5	14.9
extra light red	turističen	16	183	11.0	23.6
light red	visok	8	353	3.1	23.5
extra light red	koprski	12	110	11.8	23.2
light red	mesten	11	190	7.2	21.7
light green	siv	23	14	20.9	8.1

Appendix 2: Thesaurus – lemma “kriza”

konflikt	0.319	spor 0.273 spopad 0.267 vojna 0.266 nasilje 0.202 boj 0.17
problem	0.303	težava 0.288 razmera 0.279 situacija 0.264 dogajanje 0.228 dogodek 0.226 stanje 0 problematika 0.18 potreba 0.172 stvar 0.172
zaplet	0.25	katastrofa 0.243 padec 0.198 zlom 0.185 razpad 0.179 izbruh 0.171 tragedija 0.17
izguba	0.24	posledica 0.21 pomanjkanje 0.209 nevarnost 0.188 pritisk 0.182 vpliv 0.176 učinek
sprememba	0.229	politika 0.196 proces 0.189 razvoj 0.184 sila 0.182 gibanje 0.182 odnos 0.177
stiska	0.223	recesija 0.211 revščina 0.178
afera	0.221	
bolezen	0.207	nesreča 0.196 bolečina 0.169
revolucija	0.205	reforma 0.203 napad 0.196 volitve 0.178 poseg 0.176 akcija 0.171
nemir	0.193	napetost 0.19
obdobje	0.187	
uspeh	0.182	poraz 0.175