Compatible Sketch Grammars for Comparable Corpora

Vladimír Benko
Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics
Comenius University in Bratislava, UNESCO Chair in Translation Studies
vladob@juls.savba.sk

Abstract

Our paper describes an on-going experiment aimed at creating a family of billion-token web corpora that could to a large extent deserve the designation "comparable": corpora are of the same size, data gathered by crawling the web at (approximately) the same time, containing similar web-specific domains, genres and registers, further pre-processed, filtered and deduplicated by the same tools, morphologically annotated by (possibly) the same tagger and made available via Sketch Engine. To overcome the problem of great differences in the existing sketch grammars for the respective languages, a set of "compatible" sketch grammars have been written that will aid contrastive linguistic research and bilingual lexicographic projects. The sketch grammars use a uniform set of rules for all word categories (parts of speech) and the resulting set of tables is displayed in a fixed order in all languages.
Keywords: comparable web corpora; sketch grammars; bilingual lexicography

1 Introduction

Ten years after its introduction to the lexicographic community at the Lorient Euralex Congress (Kilgarriff et al., 2004), Sketch Engine (*SkE*) has become a standard tool in numerous lexicographic projects, as well as in various areas of corpus-based linguistic research. Sketch grammars for corpora in many languages have been written (cf. References). Recently published open-source tools for efficient web crawling (Suchomel a Pomikálek, 2012) stimulate the building of very large web corpora, the analysis of which is hardly imaginable without a powerful summarisation machine such as *SkE*. Newly implemented *SkE* support for parallel and comparable corpora (Kovář, 2013) facilitate its use in the area of bilingual lexicography and contrastive linguistic research.

In bilingual and multilingual linguistic work with *SkE*, we often encounter the problem of sketch grammars defining the collocational profiles of a headword and its translation equivalent for the respective languages. Those sketch grammars have often been created for different purposes, having in mind different user requirements, with resulting word sketches being rather disparate, making its use for contrastive linguistics problematic. Our paper suggests an alternative approach to the creation of sketch grammars, within the framework of which the respective grammars can be made compatible for (almost) all languages.

2 The Aranea project

2.1 Why new corpora?

Besides our interest in testing the new corpus-building tools, the motive for starting a new corpus project was the lack of suitable corpora that could be used by students of foreign languages and translation studies at our university. The existing web corpora families that are available for download do not cover all the languages needed. As for corpora stored at the *SkE* web site[1], they (1) are not available for download, (2) are mostly too large for classroom use[2], and (3) have too different sketch grammars, which makes them difficult to use in a mixed-language classroom.

We expect that a set of corpora for several languages of equal size and built by a standardized methodology can not only be used for teaching purposes, but also in other areas of linguistic research (contrastive studies) and in lexicography (both mono- and bilingual).

2.2 The names

For our corpora, we have decided to use "language-neutral" Latin names denoting the language of the texts and the corpus size. The whole corpus family is called *Aranea*, and the respective members bear the appropriate language name, e.g. *Araneum Anglicum*, *Araneum Francogallicum*, *Araneum Russicum* for English, French, and Russian, respectively, etc.

Each corpus exists in several editions, differing by their sizes. The basic (medium-sized) version, *Maius* ("greater"), contains approximately 1.2 billion tokens (i.e., over 1 billion words). This size can be reached relatively quickly for all participating languages, and for the "large" ones with plenty of web data available, it usually takes just one or two days of download time. The 10% random sample of *Maius*, called *Minus* ("smaller"), is to be used for teaching purposes (e.g. for lessons in the framework of Corpus Linguistics programmes for students of foreign languages and translation studies). A 1% sample, *Minimum* ("minimal"), is not intended to be used directly by the end users, and is utilized in debugging the processing pipelines and tuning the sketch grammars. And lastly, the largest *Maximum* ("maximal") edition will contain as much data as can be downloaded from the web for the particular language, and its size is mostly determined by the configuration of the server.

1 http://www.sketchengine.co.uk/

2 According to our experience, the ideal corpus for teaching corpus linguistics is about BNC-sized, i.e. it contains some 100 million tokens. As it is not easy to prevent students from invoking search operations taking several minutes to evaluate, billion-plus token corpora proved to be quite unsuitable for teaching purposes.

2.3 Web crawling

The source data acquisition is being performed by means of *SpiderLing*[3], a web crawler optimized for collecting textual data from the web. The system contains an integrated character encoding (*chared.py*) and language recognition (*trigrams.py*) module, as well as a tool for boilerplate removal (jusText). The input seed URLs (some 1,000 for each language) have initially been harvested by BootCAT[4] (Baroni and Bernardini; 2004).

Several input parameters of the crawling process are to be set by the user, most notably the language name, a file containing sample text in the respective language (to produce a model for language recognition), a language similarity threshold (a value between 0 and 1; default 0.5), the number of parallel processes, and the crawling time.

In our processing, we usually crawled the web in 24-hour slots (the process could then be re-started) with all other values set to defaults. The only exception was crawling for Slovak and Czech, where we used 7-day slots, as the process was much slower here. The language similarity threshold also had to be changed in case of Slovak and Czech. As these languages are fairly similar, the trigram method did not seem to be able to distinguish between them sufficiently. We have therefore increased the similarity threshold value to 0.65 (saving many "good" documents, and causing many "wrong" ones to pass the filter) and removed the unwanted texts by subsequent filtration based on character frequencies .

2.4 Post-download processing

Besides the basic filtration aimed to remove texts with incorrect or misinterpreted character encoding, missing diacritics and texts with non-standard proportion of punctuation and uppercase characters, the main processing operation in this phase is tokenization. As the tokenization strategy has to be compatible with that of the corpus used to train the tagger, we decided to use the tokenizers supplied with Tree Tagger and TaKIPI for the respective languages. In the future, we would like to make use of the *unitok.py* universal tokenizing program developed at Masaryk University in Brno (Jakubíček; 2014).

2.5 Deduplication

The whole procedure (Benko; 2013) consists of three stages. The first stage detects near-duplicate documents by means of the Onion (Pomikálek; 2012) utility (similarity threshold 0.95), and the duplicate documents are deleted. The second stage deduplicates the remaining text at the paragraph level using the same procedure and settings. The tokens of the duplicate paragraphs, however, are not deleted but rather they are marked to make them "invisible" during corpus searches, while they can be displayed

3 http://nlp.fi.muni.cz/trac/spiderling
4 http://bootcat.sslmit.unibo.it/

as context at the boundary of non-duplicate and duplicate text. In the last stage, we make use of our own tool based on the fingerprint method (with ignoring punctuation, special graphics characters and digits) to deduplicate the text at the sentence level. The tokens of duplicate sentences are marked similarly to the previous stage. This last step can "clean up" the duplicities among the short segments that fail to be detected as duplicates by Onion.

As deduplication is beyond the scope of our paper, we only mention here that the process has typically removed some 20–45% of tokens in the *Maius* versions of our corpora

2.6 Morpho-syntactic annotation

For languages covered by the parameter files of Tree Tagger (Schmid; 1994), this tagger has been used to annotate the respective corpora. For Polish, the TaKIPI (Piasecki; 2007), and for Czech, the Morče (Hajič; 2004) taggers were used, respectively. The question of tools for tagging Hungarian and Ukrainian data has not been resolved yet.

2.7 Tagging-related filtration

To improve the precision of tag assignments, a series of pre- and post-tagging filters have been developed that fix issues introduced by Unicode encoding of the source text[5]. The filtration fixes known tagger issues for the respective languages, namely the misassigned tags for many punctuation and special graphic characters (that are often tagged as nouns, adjectives, or abbreviations, and sometimes even as verbs with subcategories). For some languages, an additional tag with masked subcategories for gender and number is created, that is later used by some rules within the respective sketch grammars.

2.8 Current state of the project

At present, eight language versions of the *Aranea* corpus family have been created, containing both *Maius* and *Minus* editions as follows (in chronological order): *Araneum Russicum* (Russian), *Araneum Francogallicum* (French), *Araneum Germanicum* (German), *Araneum Hispanicum* (Spanish), *Araneum Polonicum* (Polish), *Araneum Anglicum* (English), *Araneum Nederlandicum* (Dutch), and *Araneum Slovacum* (Slovak).

5 As an example we can point out the problem of the "apostrophe" character in French texts. As much as 8 different Unicode characters representing the apostrophe (with just two of them being "canonical") can be found in the texts collected from the web. As the Tree Tagger French parameter file originated in the pre-Unicode era, even one of the two "canonical" representations would not be processed (i.e., tokenized and lemmatized) properly without special measures, and tokens like "*l'*" and "*d'*", that belong to the most frequent ones, would be mistagged.

The crawling has also been done for *Araneum Bohemicum* (Czech). This data is now being pre-processed to be ready for annotation that will be performed by the Institute of Theoretical and Computational Linguistics at the Faculty of Arts of Charles University in Prague.[6]

The first stage of our project will be completed by *Araneum Hungaricum* (Hungarian), *Araneum Italicum* (Italian), and *Araneum Ukrainicum* (Ukrainian). With the exception of the last mentioned, we expect to complete the whole venture by the end of 2014.

For all of the languages mentioned, sketch grammars have been written and at least two rounds of testing have been performed for each corpus. The procedure involved is described in the following section.

3 Sketch grammars

A sketch grammar[7] is a set of rules based on the CQL (Corpus Query Language[8]) used by the Sketch Engine to generate the respective collocation profiles ("word sketches") for all lexical units (lemmas) in a corpus. The word sketches are pre-computed in advance, which makes the system user-friendly and very fast.

A sketch grammar rule consists of (1) an optional comment indicated by hash "#" character, (2) the rule type marked by an asterisk "*", (3) the rule name preceded by the equal sign "=", and (4) a list of CQL expressions. For example, a rule describing the relationship between two nouns (in English using the Penn Treebank tagset) might look as follows:

```
# noun followed by another noun
*DUAL
=modifier/modified
        2: [tag="NN.*"] 1: [tag="NN.*"]
```

The "1:" label denotes the "keyword", i.e. the lemma the word sketch is created for, and the "2:" label marks the lemma of the collocate. The "*DUAL" keyword indicates that the rule is to be used twice, the second time with swapped labels, i.e. exchanging the positions of the keyword and the collocate. The text following the slash "/" character will be used as a name for the second use of the rule.

In reality, the rules usually look slightly more complex to indicate that "intermediate" words may be present between a keyword and a collocate, or in the vicinity of them.

6 Besides Ukrainian, Czech is the only language within the *Aranea* project with no free tagging tool available.
7 https://www.sketchengine.co.uk/documentation/wiki/SkE/GrammarWriting
8 https://www.sketchengine.co.uk/documentation/wiki/SkE/CorpusQuerying

3.1 What's in a name

Unlike Juliet Capulet[9], we believe that the name is often really important, and the sketch grammar rule name is exactly such a case. On one hand, it is the only component of the sketch grammar that is not predetermined, and thus can be "virtually anything". On the other hand, the name is the only clue for the user about the contents of the respective word sketch tables, and therefore should be as informative as possible. It has, however, to be very short as the name is displayed in the heading of the respective word sketch table within a only a limited space available. Rule names longer than 10–12 characters would increase the table widths, and the resulting word sketches could possibly not fit the screen.

Most sketch grammars used for corpora available at the *SkE* site follow the naming conventions introduced by A. Kilgarriff in the first English and French sketch grammars. These rule names are motivated syntactically, i.e. they denote the syntactic function of the collocate, with that of the keyword being implied. For example the rule name:

```
=modifier/modified
```

is representing two rule names with readings as follows: "collocate is a modifier of the keyword", and "collocate is modified by the keyword", respectively.

The syntactically motivated rules are transparent and user-friendly for description of basic relationships between subjects, object, modifiers/attributes, and verbs/predicates, but in more complex cases this strategy is not easily applicable. The nature of the problems can be observed in the Czech sketch grammar written by P. Smrž (Kilgarriff et al.; 2004). Some examples of rule names are as follows:

```
is_subj_of/has_subj
is_obj4_of/has_obj4
prec_prep
gen1/gen2
```

As it can be seen, it is not really easy for the user the figure out "who is who" in the keyword – collocate – syntactic function "puzzle". Moreover, rule names like "prec_verb" do not denote any syntactic functions but rather just describe collocational relationships.

There are two notable deviations from the "traditional" rule name conventions in the sketch grammars. In the grammar for the Slovenian FidaPLUS corpus[10], S. Krek (Krek; 2006) uses rule names containing (among other features) Slovenian "case questions". For example, the."*koga-česa*" name means

9 Juliet: *"What's in a name? that which we call a rose / By any other name would smell as sweet"* (William Shakespeare: Romeo and Juliet, Act II, Scene 2).

10 http://www.sketchengine.co.uk/documentation/wiki/Corpora/FidaPLUS

that only collocates of the keyword that are in genitive case are displayed, with the syntactic function of the collocate being implied.

The second notable exception is the sketch grammar written by P. Whilelock (2010) for the Oxford English Corpus[11] (OEC) where the rule names not only name the syntactic function, but also the PoS of the keyword and the collocate and their mutual position within the collocation. For example, the "V* ADJ" rule name stands for verb modified by an adjective, with askerisk indicating the keyword.

3.2 Sketch grammar for Slovak corpora

In our Institute, the *SkE* has been extensively used since autumn 2007 with several Slovak and Czech corpora. These corpora serve as a source of lexical evidence for our monolingual and bilingual lexicographic projects, as well as for other linguistic research activities.

The sketch grammar used in our *SkE* installation has been optimized for a lexicographic use, and differs from most "traditional" grammars for corpora stored at the *SkE* web site in several aspects:

- The rule names are not motivated syntactically (i.e., they do not indicate the syntactic relationship between the keyword and the collocate) but rather collocationally
- The right-hand or left-hand position of the collocate towards the keyword is indicated explicitly in the rule name
- The keyword's PoS in the rule is not specified, i.e., it covers any PoS
- Recall is preferred over precision
- The number of rules and the order of resulting tables is fixed
- The object names within the rules are governed by the following rules:
- The keyword is denoted by the X symbol
- The keyword's grammatical attributes (mostly in unary rules) are indicated by lowercase abbreviation, e.g., gen(X) indicates the genitive case of keyword
- The collocate's PoS is indicated by an abbreviation with a leading capital letter, e.g., Aj X indicates a left-hand adjective collocate
- Y indicates a collocate that is from any PoS category
- Z indicates a collocate from any PoS category not covered by the other "explicit" rules

3.3 Rule name summary

The core of our grammar consists of rules covering four basic autosemantic word classes. Taking into account our experience with early versions of the grammar, the rules for verbs (Vb X/X Vb) and adverbs (Av X/X Av) do not distinguish the left and right position of the respective collocate.

11 http://www.sketchengine.co.uk/documentation/wiki/Corpora/OEC

For nouns, two separate rules take into account the position of the collocate (Sb X, X Sb). Similar situations can be found with adjectives (Aj X; X Aj), prepositions (Pp X; X Pp) and for immediate autosemantic collocates (Y X; X Y). The "catch all" rules for the remaining word classes (Z X; X Z) cover mostly numerals and pronouns, as well as some synsemantic word classes.

The remaining two binary (symmetric) rules map the relationship of coordination, i.e., the situation where a keyword and a collocate with compatible morphological tags are separated by a comma (X/Y , X/Y) or a conjunction (X/Y Cj X/Y).

The four trinary rules cover relationships among a keyword, collocate, and preposition in different positions (Pp Y X, Pp X Y, Y Pp X, and X Pp Y).

Our set of rules is complemented by unary rules showing the frequency distribution of the keyword's forms according to grammatical categories and subcategories..

The compatible grammars

In creating sketch grammars for a group of languages, it is convenient not to use the "native" tagsets for the respective languages, but rather to use a common symbolic notation. This can be done, e.g., by means of a macro processor (such as m4). We have, however, decided to adopt a different approach and to create a simple universal tagset ("*Araneum Universal Tagset*" – AUT) similar to that of the *Universal PoS Tagset*[12] (UPT; Petrov et al., 2011), and to map all the respective tagsets into this tagset at the source vertical data level, i.e. to create a new layer of annotation. The *AUT* contains 11 tags for "traditional" part of speech categories, 7 additional tags for other elements, and one tag to indicate errors in the mapping process.

aTag	PoS
Dt	determiner/article
Nn	noun
Aj	adjective
Pn	pronoun
Nm	numeral
Vb	verb
Av	adverb

aTag	PoS
Pp	preposition
Cj	conjunction
Ij	interjection
Pt	particle
Ab	abbreviation/acronym
Sy	symbol
Nb	number

aTag	PoS
Xx	other (content word)
Xy	other other (function word)
Yy	unknown/foreign/alien
Zz	punctuation
Er	mapping error

Table 1: Araneum Universal Tagset (*AUT*).

The compatible sketch grammar using *AUT* consists of three sections. The first part (*AUT*-based) contains unary rules showing PoS category distribution for a particular lemma. The second part is

12 The AUT PoS tags for the eleven „traditional" word classes directly correspond with those of UPT, with the difference being just in the names as we wanted to keep the names of the PoS categories identical with those used in the sketch grammar rule names introduced before the UPT tagset has been published. The additional 7 categories accommodate information provided by the respective "native" tagsets that is being ignored by UPT. For example, the "Xx" (other: content word) tag is assigned to participles in Slovak that have a category of their own in the SNK Slovak tagset.

tagset-dependent and contains unary rules showing PoS subcategories provided by the respective tagset. Due to differences in the depth of the morpho-syntactic annotation, the number of subcategories varies among the languages. With verbs, e.g., we have just 5 subcategories for Spanish, while more than 20 for Polish. The final third section (*AUT*-based) covers the collocational relationships of the respective keyword by means of binary, symmetric and trinary rules.

The compatible sketch grammar is basically identical for all the languages with one important exception: the number of intermediate tokens between a keyword and a collocate is increased by one for languages having articles in their language system.

4 Discussion and conclusion

A collocationally-based sketch grammar has (against a traditional one) several advantages. It can symmetrically cover all relationships between keywords and collocates of all word classes (parts of speech). As the PoS category is not tested for the keyword, a word sketch can be created even in cases of incorrectly assigned tags. If the same (compatible) sketch grammar is used with corpora for two or more languages, the resulting word sketches can be conveniently used in contrastive linguistic research, as well as within bilingual lexicographic projects.

The disadvantage of our approach is that not all tables for some words represent linguistically relevant relationships, and they may contain a lot of noise. We believe, however, that having a fixed number of tables gives the user a clear overview, and he or she can easily ignore the irrelevant data.

In the Appendix, we present the word sketches for the lemma “without” created by means of compatible sketch grammars from four Aranea web corpora..

5 Further work

In the near future, we plan to carry out activities within several tracks. Firstly, we would like to improve the quality of the Aranea corpus data itself (by means of better filtration, normalization and deduplication), as well as its morpho-syntactic annotation by means of long-term evaluation of the resulting word sketches. Secondly, we want to include new languages into our Aranea corpus family and to write the respective corpus grammars, at least for the languages taught at Slovak universities. And finally, we plan to tune the global parameters of the compatible sketch grammars, as well as provide language-specific improvements so that the bilingual word sketches provide more relevant results.

6 References

Baroni, M. – Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In: *Proceedings of LREC'2004*. Lisbon: ELRA ,2004.

Benko, V. (2013). Data Deduplication in Slovak Corpora. In: *Slovko 2013: Natural Language Processing, Corpus Linguistics, E-learning*. Katarína Gajdošová (Ed.), Adriána Žáková. Lüdenscheid: RAM-Verlag, 2013, pp. 27–39.

Hong, J. F. – Huang, C. R. (2007). Using Chinese Gigaword Corpus and Chinese Word Sketch in linguistic research. URL: http://www.ling.sinica.edu.tw/eip/FILES/publish/2007.7.18.93102662.8243902.pdf.

Jakubíček, M. (2014). Personal communication.

Kilgarriff, A. et al. (2004). The Sketch Engine. In: G. Williams and S.Vessier (eds.), Proceedings of the eleventh EURALEX International Congress EURALEX 2004 Lorient, France, July 6–10, 2004. Lorient: Université de Bretagne-Sud, pp. 105–116.

Khokhlova, M. (2010). Building Russian Word Sketches as Models of Phrases. In: *Proc. EURALEX 2010, Leeuwarden,* July 2010.

Kovář, V. (2013). New features in the Sketch Engine interface. Part 1. In: *SKEW-4 Workshop*, Tallinn, October 2013. URL: https://www.sketchengine.co.uk/documentation/raw-attachment/wiki/SKEW-4/Program/ske_interface_part1.pdf

Krek, S. – Kilgarriff, A. (2006). Slovene Word Sketches. In: *Proceedings of 5th Slovenian and 1st international Language Technologies Conference 2006*. October 9th – 10th 2006. Jožef Stefan Institute, Ljubljana, Slovenia

Macoveiciuc, M. – Kilgarriff, A. (2010). The RoWaC Corpus and Romanian Word Sketches. In: *Multilinguality and Interoperability in Language Processing with Emphasis on Romanian*. Edited by Dan Tufis and Corina Forascu. Romanian Academy.

Petrov, S. – Das, D. – McDonald, R. (2012). A Universal Part-of-Speech Tagset. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation* (*LREC'12*). Istanbul: ELRA, May 2012.

Piasecki, M. (2007). Polish tagger TaKIPI: Rule based construction and optimisation. TASK QUARTERLY 11, No 1–2, 151–167

Radziszewski, A. – Kilgarriff, A. – Lew, R. (2011). Polish Word Sketches In: Zygmunt Vetulani (ed.) Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 5th Language & Technology Conference. Poznań : Fundacja Uniwersytetu im. A. Mickiewicza.

Schmid, H. (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Srdanović, E. I. – Erjavec. T. – Kilgarriff, A. (2008). A web corpus and word-sketches for Japanese. In: *Journal of Natural Language Processing* 15/2, 137– 159. (reprinted in Information and Media Technologies 3/3, 2008, 529– 551)

Suchomel, V. – Pomikálek, J. (2012). Efficient Web Crawling for Large Text Corpora. In: *7th Web as Corpus Workshop* (*WAC-7*), Lyon, 2012.

Tiberius, C. – Kilgarriff, A. (2009). The Sketch Engine for Dutch with the ANW corpus. In: *Fons Verbhorum, Festschrift for Fons Moerdijk*. Instituut voor Nederlandse Lexicologie, Leiden, The Netherlands.

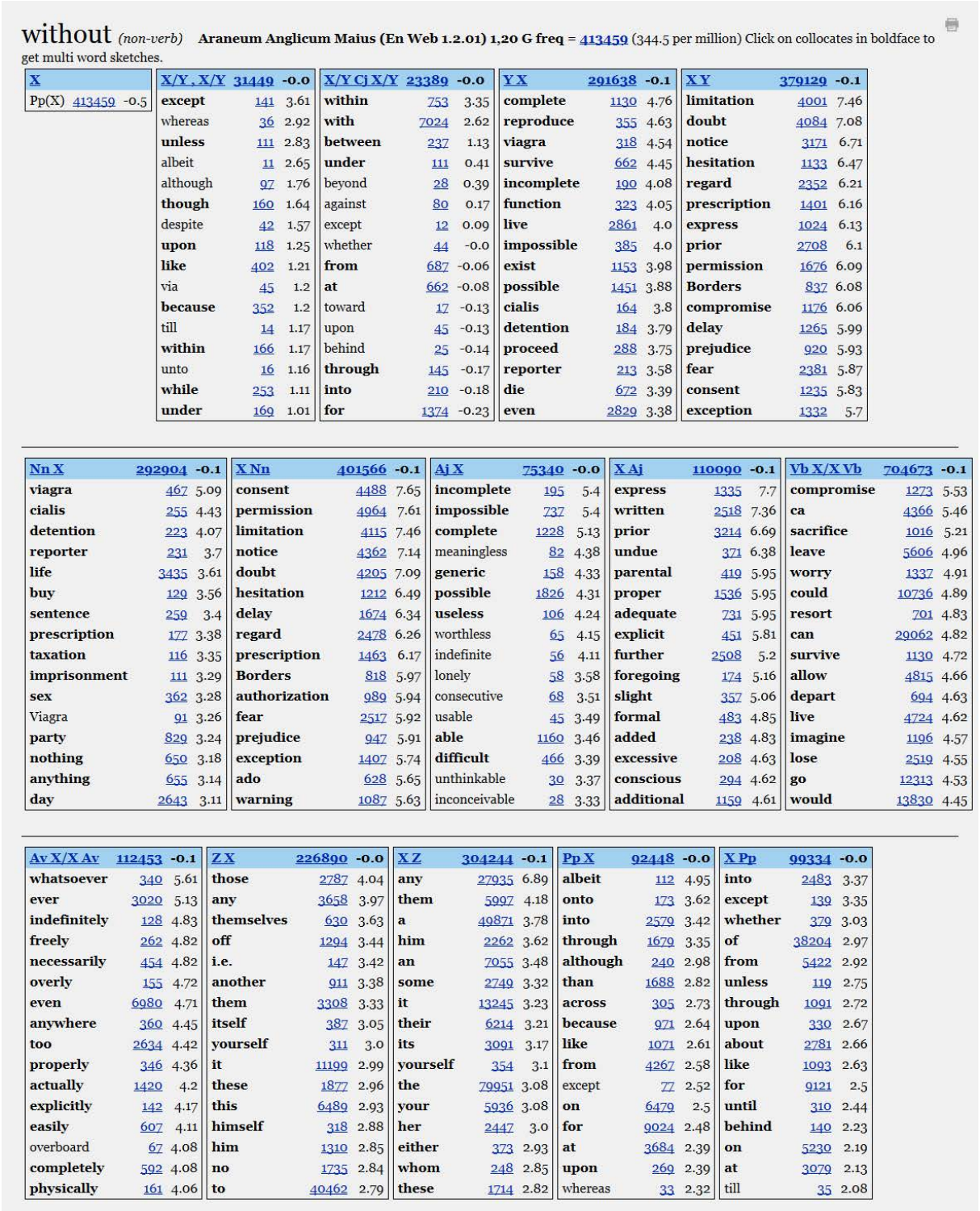
Whitelock, P. (2010). Personal communication.

Appendix

To demonstrate the compatible word sketches, we present screen shots for the preposition "without" in four languages (English, French, German, and Russian). Prepositions belong to word classes that are usually either not covered by the respective traditional sketch grammars at all, or that produce a limited number of output word sketch tables only.

For all languages involved, we can observe the typical binary collocations with noun and verbs. The collocations with adjectives usually form a multi-word expression that is not fully displayed in the word sketches, but many of those can be easily recognized even without going into the actual concordances.

Note: Due to the longer adjectives in Russian, the interesting table with verbal collocates did not fit onto the screen.



without *(non-verb)* **Araneum Anglicum Maius (En Web 1.2.01) 1,20 G freq = 413459** (344.5 per million) Click on collocates in boldface to get multi word sketches.

X	
Pp(X) 413459	-0.5

X/Y , X/Y	31449	-0.0
except	141	3.61
whereas	36	2.92
unless	111	2.83
albeit	11	2.65
although	97	1.76
though	160	1.64
despite	42	1.57
upon	118	1.25
like	402	1.21
via	45	1.2
because	352	1.2
till	14	1.17
within	166	1.17
unto	16	1.16
while	253	1.11
under	169	1.01

X/Y Cj X/Y	23389	-0.0
within	753	3.35
with	7024	2.62
between	237	1.13
under	111	0.41
beyond	28	0.39
against	80	0.17
except	12	0.09
whether	44	-0.0
from	687	-0.06
at	662	-0.08
toward	17	-0.13
upon	45	-0.13
behind	25	-0.14
through	145	-0.17
into	210	-0.18
for	1374	-0.23

Y X	291638	-0.1
complete	1130	4.76
reproduce	355	4.63
viagra	318	4.54
survive	662	4.45
incomplete	190	4.08
function	323	4.05
live	2861	4.0
impossible	385	4.0
exist	1153	3.98
possible	1451	3.88
cialis	164	3.8
detention	184	3.79
proceed	288	3.75
reporter	213	3.58
die	672	3.39
even	2829	3.38

X Y	379129	-0.1
limitation	4001	7.46
doubt	4084	7.08
notice	3171	6.71
hesitation	1133	6.47
regard	2352	6.21
prescription	1401	6.16
express	1024	6.13
prior	2708	6.1
permission	1676	6.09
Borders	837	6.08
compromise	1176	6.06
delay	1265	5.99
prejudice	920	5.93
fear	2381	5.87
consent	1235	5.83
exception	1332	5.7

Nn X	292904	-0.1
viagra	467	5.09
cialis	255	4.43
detention	223	4.07
reporter	231	3.7
life	3435	3.61
buy	129	3.56
sentence	259	3.4
prescription	177	3.38
taxation	116	3.35
imprisonment	111	3.29
sex	362	3.28
Viagra	91	3.26
party	829	3.24
nothing	650	3.18
anything	655	3.14
day	2643	3.11

X Nn	401566	-0.1
consent	4488	7.65
permission	4964	7.61
limitation	4115	7.46
notice	4362	7.14
doubt	4205	7.09
hesitation	1212	6.49
delay	1674	6.34
regard	2478	6.26
prescription	1463	6.17
Borders	818	5.97
authorization	989	5.94
fear	2517	5.92
prejudice	947	5.91
exception	1407	5.74
ado	628	5.65
warning	1087	5.63

Aj X	75340	-0.0
incomplete	195	5.4
impossible	737	5.4
complete	1228	5.13
meaningless	82	4.38
generic	158	4.33
possible	1826	4.31
useless	106	4.24
worthless	65	4.15
indefinite	56	4.11
lonely	58	3.58
consecutive	68	3.51
usable	45	3.49
able	1160	3.46
difficult	466	3.39
unthinkable	30	3.37
inconceivable	28	3.33

X Aj	110090	-0.1
express	1335	7.7
written	2518	7.36
prior	3214	6.69
undue	371	6.38
parental	419	5.95
proper	1536	5.95
adequate	731	5.95
explicit	451	5.81
further	2508	5.2
foregoing	174	5.16
slight	357	5.06
formal	483	4.85
added	238	4.83
excessive	208	4.63
conscious	294	4.62
additional	1159	4.61

Vb X/X Vb	704673	-0.1
compromise	1273	5.53
ca	4366	5.46
sacrifice	1016	5.21
leave	5606	4.96
worry	1337	4.91
could	10736	4.89
resort	701	4.83
can	29062	4.82
survive	1130	4.72
allow	4815	4.66
depart	694	4.63
live	4724	4.62
imagine	1196	4.57
lose	2519	4.55
go	12313	4.53
would	13830	4.45

Av X/X Av	112453	-0.1
whatsoever	340	5.61
ever	3020	5.13
indefinitely	128	4.83
freely	262	4.82
necessarily	454	4.82
overly	155	4.72
even	6980	4.71
anywhere	360	4.45
too	2634	4.42
properly	346	4.36
actually	1420	4.2
explicitly	142	4.17
easily	607	4.11
overboard	67	4.08
completely	592	4.08
physically	161	4.06

Z X	226890	-0.0
those	2787	4.04
any	3658	3.97
themselves	630	3.63
off	1294	3.44
i.e.	147	3.42
another	911	3.38
them	3308	3.33
itself	387	3.05
yourself	311	3.0
it	11199	2.99
these	1877	2.96
this	6489	2.93
himself	318	2.88
him	1310	2.85
no	1735	2.84
to	40462	2.79

X Z	304244	-0.1
any	27935	6.89
them	5997	4.18
a	49871	3.78
him	2262	3.62
an	7055	3.48
some	2749	3.32
it	13245	3.23
their	6214	3.21
its	3091	3.17
yourself	354	3.1
the	79951	3.08
your	5936	3.08
her	2447	3.0
either	373	2.93
whom	248	2.85
these	1714	2.82

Pp X	92448	-0.0
albeit	112	4.95
onto	173	3.62
into	2579	3.42
through	1679	3.35
although	240	2.98
than	1688	2.82
across	305	2.73
because	971	2.64
like	1071	2.61
from	4267	2.58
except	77	2.52
on	6479	2.5
for	9024	2.48
at	3684	2.39
upon	269	2.39
whereas	33	2.32

X Pp	99334	-0.0
into	2483	3.37
except	139	3.35
whether	379	3.03
of	38204	2.97
from	5422	2.92
unless	119	2.75
through	1091	2.72
upon	330	2.67
about	2781	2.66
like	1093	2.63
for	9121	2.5
until	310	2.44
behind	140	2.23
on	5230	2.19
at	3079	2.13
till	35	2.08

ohne **Araneum Germanicum Maius (De Web 1.2.01) 1,20 G freq** = 592115 (493.4 per million) Click on collocates in boldface to get multi word sketches.

X	
Pp(X)	505192 3.0
Cj(X)	86923 0.9

X/Y , X/Y	34121	0.2
dass	1892	4.32
ob	270	3.32
oder	1064	3.25
d.h.	21	3.05
wie	659	2.85
und	5054	2.84
einschließlich	39	2.69
bzw.	54	2.23
inklusive	30	2.2
außer	45	2.17
weil	61	2.1
als	464	2.06
infolge	12	1.72
indem	12	1.6
aufs	20	1.43
sondern	55	1.37

X/Y Cj X/Y	41201	0.2
einschließlich	57	3.19
seitens	40	2.8
mit	8640	2.61
inklusive	34	2.33
zwischen	382	1.97
trotz	71	1.8
mittels	40	1.8
binnen	12	1.79
außer	30	1.55
samt	12	1.54
aufs	22	1.53
außerhalb	39	1.51
wider	11	1.47
ans	18	1.4
ob	71	1.38
gegen	269	1.36

Y X	323276	0.3
Kredit	2342	6.4
jederzeit	1222	5.36
Reporter	414	5.2
gänzlich	475	5.11
ganz	8310	4.75
Fahren	306	4.63
völlig	929	4.54
Fass	236	4.34
Abnehmen	236	4.32
Girokonto	288	4.28
Handy	539	4.26
Kreditkarte	348	4.24
Rechnung	486	4.13
Leben	2437	4.12
allerdings	1516	4.09
Pfanne	226	4.03

X Y	483395	0.4
vorherig	4624	7.62
Schufa	3481	7.57
Zweifel	3621	7.26
Rücksicht	2869	7.24
Weiteres	1810	6.82
Einschränkung	2002	6.49
Angabe	4455	6.46
gesondert	1607	6.35
Zustimmung	2145	6.3
Abzug	1372	6.19
Gewähr	1324	6.19
Problem	6810	6.14
Umweg	1192	6.11
Unterbrechung	1176	6.11
ausdrücklich	2164	6.1
Behinderung	1596	5.93

Nn X	311082	0.3
Kredit	2584	6.56
Publikation	1004	5.87
Reporter	415	5.26
Fahren	337	4.82
Rechnung	739	4.76
Handy	708	4.67
Girokonto	340	4.55
Kreditkarte	402	4.48
Abnehmen	247	4.43
Fass	239	4.41
Leben	2895	4.37
Angebot	1875	4.2
Tarif	439	4.19
Pfanne	242	4.17
Abmahnung	267	4.13
Baufinanzierung	197	4.05

X Nn	606564	0.5
Zustimmung	5620	7.5
Schufa	3453	7.29
Zweifel	4051	7.21
Ankündigung	3067	7.07
Rücksicht	3041	7.07
Aufwand	3653	6.93
Einschränkung	2510	6.59
Genehmigung	2522	6.55
Weiteres	1818	6.52
Angabe	4874	6.49
Grund|Gründen	3484	6.45
Einwilligung	1781	6.31
Problem	7956	6.31
Abzug	1758	6.29
Umweg	1353	6.01
Gewähr	1384	5.99

Aj X	89134	0.2
gedruckt	973	7.25
gänzlich	485	6.2
völlig	1022	4.96
anwaltlichen	83	4.73
gesamt	1581	4.62
erhoben	133	4.52
Prepaid	112	4.5
personenbezogen	248	4.5
komplett	820	4.38
selbstverständlich	433	4.38
viagra	61	4.27
vollkommen	226	4.26
undenkbar	64	4.24
kommerziell	153	4.22
verlinkten	97	4.19
berufsmäßig	52	4.19

X Aj	200309	0.5
vorherig	4938	8.37
gesondert	1704	7.27
ausdrücklich	3061	7.06
nennenswert	811	6.66
schriftlich	2599	6.55
lästig	639	6.18
erkennbar	856	6.09
störend	448	5.8
weit	11902	5.58
zusätzlich	3407	5.58
möglich	4803	5.48
ersichtlich	345	5.34
fremd	827	5.32
unnötig	449	5.28
aufwendig	523	5.21
finanziell	1189	5.19

Vb X/X Vb	
auskommen	
gestatten|stat	
funktioniere	
verändern	
verlaufen	
verlieren	
leben	
laufen	
zögern	
dürfen	
verlassen	
nachdenken	
bleiben	
können	
kündigen	
überstehen	

Av X/X Av	173826	0.4
jemals	724	6.31
jederzeit	1696	6.07
vorher	1598	5.93
jedoch	4386	5.15
allerdings	2735	5.01
ganz	9471	4.96
niemals	434	4.86
leider	1386	4.66
irgend	201	4.65
kaum	1261	4.63
freilich	204	4.62
also	3264	4.58
bisher	1131	4.56
auch	35534	4.56
sofort	1136	4.53
meistens	430	4.52

Z X	285861	0.3
man	8344	4.02
14	690	3.93
diese	1417	3.91
nicht	20169	3.85
wer	1280	3.83
keine	4041	3.74
daher	740	3.57
solche	1093	3.54
sie	6700	3.53
niemand	257	3.45
eine	37041	3.41
jemand	317	3.39
was	2380	3.32
er	5165	3.25
nichts	655	3.24
es	9643	3.23

X Z	356121	0.4
jegliche	4363	7.76
irgendwelche	1088	6.08
dabei	6094	5.38
irgendeine	556	4.95
allzu	476	4.74
dafür	1743	4.61
jede	5801	4.4
zu	52715	4.36
nichts	1282	4.16
keine	4669	3.94
solche	1384	3.85
jedwede	159	3.75
darauf	740	3.74
jemand	424	3.71
viele	4400	3.71
darüber	676	3.69

Pp X	67488	0.1
dank	140	3.13
pro	258	2.87
bei	5283	2.86
einschließlich	51	2.85
gegenüber	221	2.71
inklusive	49	2.7
trotz	135	2.67
nach	2766	2.58
innerhalb	209	2.49
für	7039	2.48
wegen	200	2.46
seit	561	2.46
zeit	20	2.46
von	8310	2.43
mittels	63	2.36
binnen	20	2.23

X Pp	111231	0.3
seitens	198	4.65
durchs	80	3.87
ans	124	3.79
auf	15175	3.71
außer	138	3.48
aufs	105	3.43
von	16501	3.42
durch	3718	3.34
binnen	56	3.32
über	3330	3.0
an	7594	2.92
innerhalb	277	2.84
per	225	2.79
in	22140	2.69
mit	8122	2.52
gegen	606	2.5

sans **Araneum Francogallicum Maius (Fr Web 1.2.02) 1,23 G freq** = 1023016 (829.5 per million) Click on collocates in boldface to get multi word sketches.

X		
Pp(X)	1012928	-0.6
Cj(X)	10088	-0.0

X/Y , X/Y	165883	-0.2
sauf	248	3.9
jusque	599	3.29
malgré	217	3.13
hormis	51	3.03
hors	147	3.01
avec	4598	2.94
depuis	869	2.89
chez	541	2.89
vers	626	2.88
pendant	439	2.83
sous	725	2.77
dès	262	2.63
à	18464	2.62
au	7379	2.59
devant	323	2.58
en	12377	2.58

X/Y Cj X/Y	89170	-0.1
avec	11553	4.27
hors	104	2.71
soit	77	2.17
malgré	78	1.8
parce	36	1.58
entre	519	1.56
sauf	41	1.52
en	5765	1.48
sous	284	1.45
pendant	162	1.45
contre	250	1.37
à	7766	1.37
dès	98	1.29
du	9242	1.28
jusque	137	1.22
de	23057	1.17

Y X	638024	-0.1
reporter	975	5.06
non	5532	4.86
c'est-à-dire	804	4.49
fonctionner	1169	4.48
organisme	1212	4.2
répéter	637	4.16
rester	3613	4.16
accepter	1380	4.15
presque	971	4.14
vivre	2985	4.14
consommer	669	4.14
interdire	864	4.01
mourir	943	3.98
vétérinaire	364	3.94
voiture	857	3.87
dérouler	816	3.85

X Y	878596	-0.1
doute	85421	9.99
cesse	20493	9.02
oublier	15100	7.48
fil	7755	7.07
précédent	6972	6.97
préavis	3701	6.96
autorisation	5545	6.84
autant	9143	6.81
relâche	3002	6.72
compter	12738	6.67
frontière	4720	6.61
conteste	2740	6.6
faille	3165	6.6
consentement	2514	6.28
moindre	3757	6.26
limite	4816	6.24

Nn X	526473	-0.1
reporter	920	5.16
organisme	1441	4.53
journée	1822	4.3
voiture	1075	4.28
nuit	1242	4.16
Internet	1353	4.07
réseau	2004	4.07
acceptation	367	4.02
amour	1131	3.91
jour	4405	3.88
médecin	739	3.88
licenciement	300	3.86
connexion	400	3.79
monde	3438	3.77
sexe	522	3.74
gens	1270	3.74

X Nn	969635	-0.1
doute	96148	10.11
cesse	20510	8.93
fil	7829	7.02
préavis	3832	6.89
autorisation	6013	6.88
précédent	6605	6.82
relâche	3004	6.59
faille	3200	6.49
conteste	2760	6.47
limite	5331	6.32
surprise	3308	6.29
consentement	2647	6.24
souci	3214	6.03
hésitation	2068	6.02
arrêt	3353	6.0
frontière	3229	5.98

Aj X	118195	-0.1
vétérinaire	260	5.06
impossible	736	5.01
tierce	139	4.55
modifiable	97	4.54
gratuit	747	4.34
accessible	517	4.29
possible	1898	4.26
utilisable	108	4.08
remboursable	69	3.93
sexuel	320	3.89
correct	133	3.85
partiel	173	3.8
réalisable	63	3.71
estre	81	3.7
immédiat	197	3.67
inconcevable	53	3.66

X Aj	167461	-0.1
lucratif	2851	8.2
préalable	4847	7.94
moindre	5688	7.7
expresse	1296	7.5
nul	2863	7.15
apparent	1106	6.93
frais	3179	6.59
frontière	1676	5.97
égal	1040	5.64
réel	2466	5.61
gras	748	5.6
fixe	640	5.5
excessif	420	5.44
supplémentaire	1195	5.42
valable	526	5.37
explicite	298	5.17

Vb X/X Vb	1329224	-0.1
oublier	16163	7.3
compter	14308	6.6
parler	12885	5.9
soucier	2432	5.7
tarder	2551	5.7
laisser	9961	5.6
attendre	7517	5.6
perdre	5087	5.3
pouvoir	54251	5.3
passer	13149	5.2
hésiter	2779	5.1
savoir	12005	5.1
rester	7627	5.0
vivre	6437	5.0
bouger	1713	5.0
regarder	3542	5.0

Av X/X Av	287079	-0.1
autant	10797	7.41
toutefois	3317	6.99
jamais	11860	6.7
trop	11522	6.58
forcément	1325	5.84
nécessairement	855	5.74
rien	6754	5.71
vraiment	3859	5.5
non	7636	5.4
priori	564	5.33
réellement	926	5.33
c'est-à-dire	989	5.31
exprès	387	5.23
presque	1466	5.05
préalablement	416	5.02
apparemment	419	4.92

Z X	526005	-0.1
ceci	1004	4.96
cela	6226	4.58
la|le	9272	4.47
la	2222	4.32
se	32382	4.28
un	77428	4.26
on	10142	4.24
elle	9775	4.24
il	25892	4.1
votre	5454	4.07
vous	12695	4.07
tu	2004	3.99
ce	38418	3.91
y	5689	3.85
leur	5436	3.81
qui	24786	3.8

X Z	554614	-0.1
aucun	36125	8.34
quoi	3522	5.73
OGM_	963	5.49
la|le	10965	4.71
la	2817	4.65
toi	1128	4.64
eux	1697	4.46
lequel	3094	4.41
lui	6975	4.33
y	7918	4.32
leur	7251	4.22
se	30956	4.21
me	6916	4.14
son	19423	4.14
que	19955	4.11
votre	5045	3.96

Pp X	258101	-0.1
devant	691	3.61
chez	870	3.52
vers	992	3.49
du	42461	3.48
à	31897	3.41
de	107821	3.4
pendant	677	3.39
après	1187	3.35
dans	11797	3.18
pour	11453	3.06
avec	4790	2.99
sur	7429	2.92
au	9114	2.9
derrière	177	2.85
contre	726	2.83
en	14810	2.83

X Pp	349901	-0.1
jusque	1966	4.88
pour	27339	4.31
avec	9295	3.94
pendant	1000	3.89
sur	13933	3.83
à	42038	3.8
depuis	1684	3.78
de	137573	3.75
au	15714	3.68
par	11444	3.67
devant	715	3.59
dans	15547	3.58
vers	1043	3.52
contre	1189	3.51
envers	242	3.43
quant	240	3.42

без

Araneum Russicum Maius (Ru Web 1.2.02) 1,22 G freq = 749749 (616.2 per million) Click on collocates in boldface to get multi word sketches.

X		
Pp(X)	749749	-2.3
nom(X)	6477	-0.0
gen(X)	737804	-1.3
dat(X)	1640	-0.0
acc(X)	1003	-0.0
loc(X)	2266	-0.0
ins(X)	559	-0.0

X/Y , X/Y	16565	-0.2
безо	15	3.46
вне	90	3.05
кроме	230	2.46
ради	54	1.86
посредством	18	1.21
вместо	39	0.99
путем	24	0.98
помимо	24	0.91
среди	74	0.65
выше	16	0.6
возле	11	0.56
от	782	0.4
для	989	0.35
из	811	0.33
у	552	0.27
внутри	22	0.27

X/Y Cj X/Y	9119	-0.1
вне	76	2.84
у	526	0.2
помимо	12	-0.07
вместо	16	-0.28
против	44	-0.28
вокруг	14	-0.74
для	445	-0.81
со	90	-0.84
до	163	-0.86
из-за	22	-0.91
от	294	-1.01
после	78	-1.19
среди	18	-1.39
из	237	-1.44
с	296	-2.38

Y X	424569	-0.3
обойтись	12297	8.55
обходиться	5964	8.15
остаться	14808	6.92
пропасть	2590	6.68
невозможный	2621	6.53
оставить	4941	6.24
невозможно	2174	6.07
немыслимый	954	5.93
оставлять	1678	5.37
пропавшими	485	5.2
практически	2233	5.14
оставаться	3850	5.12
прожить	741	4.98
жить	4890	4.89
вообще	2127	4.79
почти	1985	4.67

X Y	620883	-0.3
исключение	9151	7.69
сомнение	8080	7.63
весть	4171	7.49
попечения	2636	7.02
преувеличение	2256	6.76
особый	9784	6.71
труд	8844	6.61
лишний	4748	6.58
учет	6474	6.55
малое	1960	6.51
предварительный	2943	6.33
малейший	2032	6.25
присмотр	1582	6.25
посредник	1786	6.21
согласие	2584	6.19
разбор	1567	6.16

Nn X	364089	-0.3
оставление	423	5.12
кредит	1386	4.88
репортер	312	4.68
дым	377	4.51
участок	1432	4.4
кофе	432	4.26
чай	565	4.17
лицо	2825	4.14
секс	396	4.02
отпуск	412	4.01
наличные	219	3.98
квартира	1366	3.95
бой	586	3.93
жизнь	5441	3.92
столбик	187	3.91
заработок	387	3.84

X Nn	844256	-0.4
сомнение	9447	7.64
исключение	9375	7.52
весть	4173	7.12
труд	10846	6.8
согласие	4377	6.69
попечения	2737	6.66
преувеличение	2603	6.56
ведомо	2464	6.49
учет	6771	6.48
разрешение	4377	6.35
ограничение	3996	6.31
усилие	4131	6.25
малое	1968	6.12
потеря	3707	6.11
ущерб	2715	6.09
вмешательство	2359	6.03

Aj X	88968	-0.2
немыслимый	1008	7.55
пропавшими	471	7.31
невозможный	2767	7.25
пропавшим	176	5.94
худой	324	5.78
мыслимый	187	5.55
предпринимательский	249	5.14
минеральный	263	4.66
неполный	156	4.63
исковый	124	4.43
ровный	205	4.27
гладкий	146	4.12
земельный	290	3.92
апелляционный	83	3.91
длительный	397	3.85
Послеоперационный	39	3.82

X Aj	194940	-0.3
малейший	2061	7.24
предварительный	3438	7.19
видимый	1687	7.12
лишний	4940	7.08
особый	10017	6.96
посторонний	1570	6.9
уважительный	761	6.6
должный	1241	6.53
невозможный	1714	6.32
дополнительный	4464	6.13
излишний	669	5.86
надлежащий	563	5.67
немыслимый	397	5.53
хирургический	475	5.48
единый	2288	5.27
специальный	3111	5.13

Vb X/
обойт
обход
остат
остав
проп
остав
прож
жить
остав
преде
смочь
мысл
позво
мочь
спра
выжи

Av X/X Av	100363	-0.2
невозможно	7192	8.32
немыслимо	439	6.99
нельзя	4436	6.65
извне	309	5.98
практически	3089	5.85
трудно	1261	5.83
сложно	797	5.79
вообще	3670	5.78
скучно	252	5.77
вовсе	983	5.53
бесплатно	507	5.37
можно	13945	5.32
почти	2600	5.25
желательно	325	4.96
возможно	826	4.93
тяжело	302	4.91

Z X	231029	-0.2
куда	1227	5.54
тут	1299	4.65
даже	4526	4.49
ж	324	4.4
никак	412	4.35
поэтому	1540	4.28
ведь	1216	4.15
теперь	1167	4.12
здесь	1347	4.05
не	40489	3.96
бы	4480	3.82
весь	8436	3.81
тоже	1295	3.73
туда	277	3.73
как-то	277	3.62
просто	1910	3.61

X Z	226728	-0.3
всякий	21391	8.49
какой-либо	10149	8.07
никуда	887	6.39
никак	1459	6.18
таковой	1174	6.03
оное	252	5.04
оный	281	4.87
чей-либо	190	4.68
какой	2738	3.89
ничто	1527	3.82
то	13929	3.79
они	13906	3.63
она	9102	3.56
он	15472	3.55
пять	539	3.45
ваш	2980	3.43

Pp X	70963	-0.1
ко	271	3.49
сверх	23	2.68
за	2747	2.45
на	12506	2.44
обо	39	2.41
до	1499	2.34
во	994	2.33
посреди	19	2.29
в	23004	2.25
через	483	2.21
при	1481	2.2
к	3292	2.16
из	2797	2.11
об	486	2.09
сквозь	29	2.09
под	609	2.06

X Pp	79547	-0.1
со	2113	3.69
над	724	3.57
для	6323	3.02
на	16431	2.83
к	4975	2.75
с	8158	2.4
от	3131	2.4
о	2583	2.28
из-за	206	2.24
об	527	2.21
свыше	40	2.21
около	223	2.2
до	1265	2.09
пред	27	2.07
между	375	2.04
в	19486	2.01