NATURAL LANGUAGE PROCESSING AND SOCIAL NETWORK ANALYSIS

STUDYING SPECIAL TEXT RUSSIAN CORPORA BY THE LEXICO-SYNTACTIC MODELS

Maria Khokhlova, Victor Zakharov

Abstract: *The paper presents the results of automatic term extraction from a special text corpus (a collection of papers on corpus linguistics) by means of statistical methods (association measures) combined with certain syntactic models. The approach undertaken in the paper is based on lexico-syntactic models that can be viewed as models of phrases for the Russian language. The Sketch Engine system represents itself a corpus tool which takes as input a corpus of any language and corresponding grammar patterns. The system gives information about a word's collocability on concrete dependency models, and generates lists of the most frequent phrases for a given word based on appropriate models. The extracted terms belong to various clusters and represent the lexical structure of the texts in question. The applied method includes statistical analysis that enables estimating paradigmatic and syntagmatic relations between lexemes based on their distribution.*

Keywords: *Corpora, distributional and statistical methods, collocations, automatic term extraction, thesaurus.*

ACM Classification Keywords*: I.2.7 Natural Language Processing*

Introduction

This research was based on using the Sketch Engine system, a corpus tool which takes as input a corpus of any language and corresponding grammar patterns and which generates word sketches for words of that language [Kilgarriff, et al., 2004; The Sketch Engine]. We have developed syntactic patterns (models of phrases or word sketches) for the Russian language based on a morphologically annotated corpus. These syntactic patterns can be viewed as lexico-syntactic models of phrases. One can understand word sketches as typical phrases determined on the one hand by syntax that restricts words' collocability in a given language and on the other hand by the probability closely related to word usage [Kilgarriff, et al., 2004; Rychly, Smrz, 2004; Mel'cuk, 1998].

Lexico-Syntactic Models for the Russian Language

Lexico-syntactic model is a structural pattern of a linguistic construction having an indication of grammatical properties of a number of lexemes (that belong to the construction) and of syntactic conditions of using the verbal expression built according to the pattern (for example, rules for agreement of morphological properties of the lexemes in phrases, e.g. in Russian adjectives agree with nouns in gender,

case, and number). This approach can be supplemented with a statistical approach that takes into account frequencies of words and their combinations [Bol'shakova et al., 2007; Mitrofanova, Zakharov, 2009].

The notion of lexico-syntactic model was applied while describing syntactic phrases for the Russian language. We developed syntactic patterns (models of phrases) for the Russian language based on a morphologically annotated corpus. This grammar was integrated to the Sketch Engine that generates on its basis word sketches reflecting the words' lexical and syntactic collocability. While writing word sketch grammar we used theses described in [Russkaya grammatika, 1980; Zolotova, 1988; Bol'shakov, Bol'shakova, 2006].

While describing the syntactic patterns inherent to the Russian language we distinguished the following models:

- coordination (=and/or);
- subjective model (N_1+V: =subject/subject_of, =passive/subj_passive, =to_be_adj/subj_to_be);
- objective model (V+N_2, V+N_3, V+N_4, V+N_5: =object2/object2_of, =object3/object3_of, =object4/object4_of, =inst_modifier/inst_modifies; V+Vinf: =post_inf/verb_post_inf; Adj_short+V: =modal_inf/modal);
- attributive model (N+N_2: =gen_modifier/gen_modifier; Adj+N =a_modifier/modifies);
- comparative model (N+Adjcomp+N_2: =comparative);
- adverbial model (=adv_modifier/adv_modifies);
- prepositional model (Prep+N, V+Prep: =prec_prep, =post_prep; N+PP, V+PP: =pp_%s, =pp_obj_%s).

Within the models there are 18 relations. Thus, the attributive model can be described by two relations that are N+N_2 and Adj+N.

The latter relation includes three cases: 1) Adj + (Adj) + N; 2) Adj + («,» + Adj) + «или» ('or') / «и» ('and') +Adj + N; 3) Adj + («,» + Adj) + N. Here are the examples from the corpus taking into account each case respectively:

- *Рассмотрим в качестве иллюстрации высказанной мысли* ***английские синонимичные прилагательные*** *(English synonymic adjectives) easy, simple (на основе COCA) (Adj + (Adj) + N).*
- *Подкорпус содержит только целые тексты, имеющие* ***метатекстовую, морфологическую и семантическую разметку*** metatextual*, morphological* and *semantic annotation) (Adj + («,» + Adj) + «или» ('or') /* ***«и» ('*and*')*** *+Adj + N).*
- *Особенностью данных топонимического корпуса является то, что несмотря на однотипность базовых единиц корпуса – топонимов, наблюдается существенная их неоднородность как в плане языковой принадлежности (***русский, финский, ижорский, водский, эстонский, шведский, немецкий языки***) (Russian, Finnish,* Ingrian, *Votic, Estonian*, Swedish*, German languages), так и в плане характеристик материалов и их носителей (карточки, карты, списки и другие источники) (*Adj + (**«,»** + Adj) + N*).*

As next stage of our research we uploaded the described grammar into the Sketch Engine to get examples of terms and phrases from the corpus.

Special Text Corpus

Specialized languages occupy a prominent place in both linguistics and the information technology. This study implies a scientific text to be treated as a special one. Special texts are rich in terminology and that calls for developing methods to automatically extract terms from a special text corpus. The corpus itself is a collection of papers on corpus linguistics published in a number of conference proceedings in Russian, totally about 343000 tokens. The automatic term extraction is based on grammatical patterns and statistics allowing to weight terms, to estimate them. The area of corpus linguistics is a rapidly developing field of linguistics with its own methodology and terms [Mitrofanova, Zakharov, 2009]. Moreover a vast majority of terms come from the English language, and sometimes there is no agreement in spelling (for example, «тэг» or «тег» for the English ‘tag’).

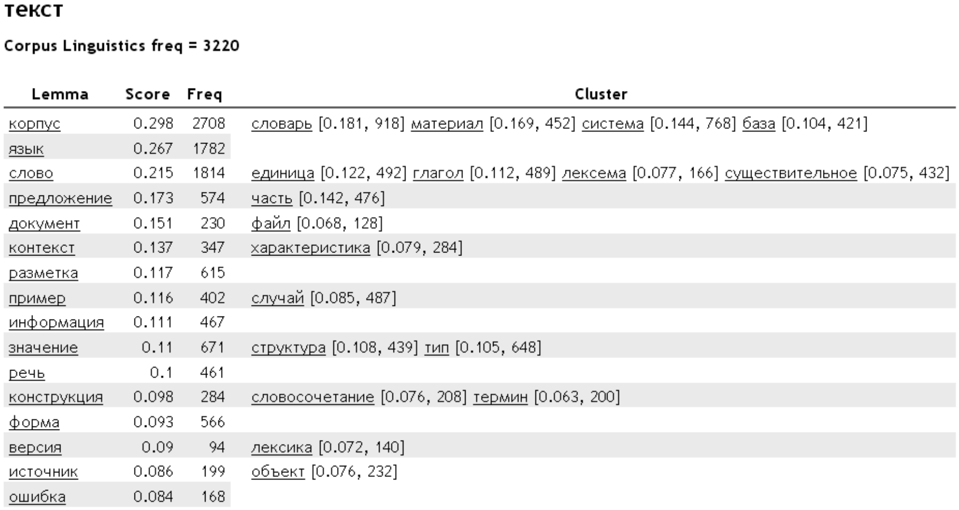
In terms of linguistics, we are talking about a plethora of units (notions) defined by the terms “lexical field”, “lexico-semantic field”, and “functional semantic field”. In modern information technology the same notions can be normally called thesaurus or ontology.

The Sketch Engine has special tools that allow to measure syntagmatic and paradigmatic relations between lexical units based on lexemes distribution and syntactical collocatibility rules: Word Sketches, Thesaurus, Clustering, and Differences. The statistical measures enable ranking the extracted terms; most of them represent set phrases and collocations.

Experiments

Below there’s a list of the 30 most frequent single-word terms (with a high index of collocability) found in the corpus: “tekst” (‘text’), “korpus” (‘corpus’), “slovo” (‘word’), “yazyk” (‘language’), “slovar’” (‘lexicon’, ‘vocabulary’), “dannyje” (‘data’), “sistema” (‘system’), “znachenije” (‘meaning’, ‘sense’), “tip” (‘type’), “razmetka” (‘tagging’, ‘annotation’), “analiz” (‘analysis’), “predlozhenije” (‘sentence’), “forma” (‘form’), “issledovanije” (‘research’), “rabota” (‘work’), “vremya” (‘tense’), “jedinitsa” (‘item’), “glagol” (‘verb’), “sluchaj” (‘case’), “chast’” (‘part’), “informatsiya” (‘information’), “sozdaniye” (‘creation’), “rech” (‘speech), “material” (‘material’), “struktura” (‘structure’), “suscestvitel’noye” (‘noun’), “baza” (‘base’), “primer” (‘example’), “zadacha” (‘task’), and “svyaz’” (‘relation’).

The output for each term is represented by a class of words that can be semantically related to it. Below in Fig. 1 one can see the results for the key word “tekst” (‘text’) selected by the Thesaurus function with enabled clustering option:



текст

Corpus Linguistics freq = 3220

Lemma	Score	Freq	Cluster
корпус	0.298	2708	словарь [0.181, 918] материал [0.169, 452] система [0.144, 768] база [0.104, 421]
язык	0.267	1782	
слово	0.215	1814	единица [0.122, 492] глагол [0.112, 489] лексема [0.077, 166] существительное [0.075, 432]
предложение	0.173	574	часть [0.142, 476]
документ	0.151	230	файл [0.068, 128]
контекст	0.137	347	характеристика [0.079, 284]
разметка	0.117	615	
пример	0.116	402	случай [0.085, 487]
информация	0.111	467	
значение	0.11	671	структура [0.108, 439] тип [0.105, 648]
речь	0.1	461	
конструкция	0.098	284	словосочетание [0.076, 208] термин [0.063, 200]
форма	0.093	566	
версия	0.09	94	лексика [0.072, 140]
источник	0.086	199	объект [0.076, 232]
ошибка	0.084	168	

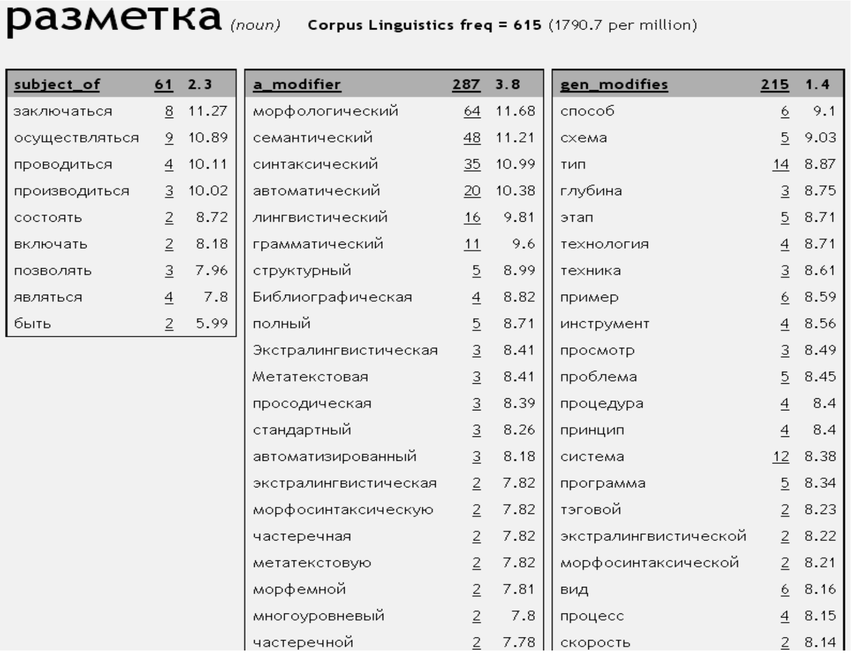
Fig. 1. Thesaurus for the word “tekst” (‘text’)

Among the presented results one can distinguish between the following clusters: 1) “source of research” — “korpus” (‘corpus’), “slovar’” (‘dictionary’), “material” (‘material’), “sistema” (‘system’), “baza” (‘basis’); 2) “object of study” — “slovo” (‘word’), “jedinitsa” (‘unit’), “глагол” (‘verb’), “leksema” (‘lexeme’) etc.

Two-word terms can be extracted from the corpus by applying the Word Sketch function. Fig. 2 shows typical collocations for some frequent lexemes (sorted by frequency) that match predefined lexico-syntactic models.

All the extracted terms were grouped according to their grammatical structure. For the term “razmetka” (‘tagging’, ‘annotation’): 1) Adj N — “morfologicheskaya razmetka” (‘morphological tagging’), “semanticheskaya razmetka” (‘semantic annotation’), “syntaksicheskaya razmetka” (‘parsing’) etc.; 2) N N_2 — “glubina razmetki” (‘depth of annotation’), “uroven’ razmetki” (‘level of annotation’) etc. For the term “glagol” (‘verb’): Adj N — “frazovyj glagol” (‘phrasal verb’), “modal’nyj glagol” (‘modal verb’), “kauzativnyj glagol” (‘causative verb’) etc.

There are two groups of collocations among the extracted terms. The former includes the terms themselves that can be added to the dictionaries, the latter is represented by high frequent collocations: “opusceniye glagola” (‘omission of verbs’), “angliyskiy glagol” (‘English verb’) etc. Both groups can be used while describing the term system of corpus linguistics as such lexis is not often reflected in dictionaries.



разметка (noun) **Corpus Linguistics freq = 615** (1790.7 per million)

subject_of	61	2.3	a_modifier	287	3.8	gen_modifies	215	1.4
заключаться	8	11.27	морфологический	64	11.68	способ	6	9.1
осуществляться	9	10.89	семантический	48	11.21	схема	5	9.03
проводиться	4	10.11	синтаксический	35	10.99	тип	14	8.87
производиться	3	10.02	автоматический	20	10.38	глубина	3	8.75
состоять	2	8.72	лингвистический	16	9.81	этап	5	8.71
включать	2	8.18	грамматический	11	9.6	технология	4	8.71
позволять	3	7.96	структурный	5	8.99	техника	3	8.61
являться	4	7.8	Библиографическая	4	8.82	пример	6	8.59
быть	2	5.99	полный	5	8.71	инструмент	4	8.56
			Экстралингвистическая	3	8.41	просмотр	3	8.49
			Метатекстовая	3	8.41	проблема	5	8.45
			просодическая	3	8.39	процедура	4	8.4
			стандартный	3	8.26	принцип	4	8.4
			автоматизированный	3	8.18	система	12	8.38
			экстралингвистическая	2	7.82	программа	5	8.34
			морфосинтаксическую	2	7.82	тэговой	2	8.23
			частеречная	2	7.82	экстралингвистической	2	8.22
			метатекстовую	2	7.82	морфосинтаксической	2	8.21
			морфемной	2	7.81	вид	6	8.16
			многоуровневый	2	7.8	процесс	4	8.15
			частеречной	2	7.78	скорость	2	8.14

Fig. 2. Word sketches for the word “razmetka” (‘tagging’)

Conclusion and Further work

The above described methodology of using Sketch Engine instruments on scientific text corpus of Russian allows to extract terminological phrases (not single word terms only), to define paradigmatic relations added to syntagmatic ones, and to quantitatively estimate the strength of semantic relations.

There is a question of corpus volume. For example, different association measures extract different collocations but here one can’t see differences between results obtained by a number of statistical measures, it means that collocates will be quite the same. This problem arises from low frequencies of words and phrases. As was pointed above we are going to work on further corpus data increase.

A number of problems arise from errors in morphological annotation as: 1) every punctuation mark has its own tag (so it should be excluded in the sketch grammar); 2) parts of compound nouns also have different lemmas that is why in sketch tables we can find only one part of such words as a collocate; 3) usual mistakes of annotation, e.g. homonyms or homographs, mistakes in assigning the correct case or number; 4) mistakes in assigning correct lemmas (it is especially the case while annotating special texts).

Further development of this mechanism of collocation extraction is closely related to writing more exact grammatical rules (that will be based on syntactically parsed corpus or even take into account semantic annotation), more corpus data etc. Most errors in the word sketches result from errors in lemmatization and POS-tagging. We are currently explore alternative tools for automatic morphological annotation. Manual morphological disambiguation can be seen as a possible solution for the problem of reducing errors of annotation. But this work is labour- and time-consuming and unfortunately can be applied only to a small part of a corpus.

Bibliography

[Kilgarriff, et al., 2004] Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. (2004). The Sketch Engine. In: *Proceedings of EURALEX-2004*, 105-116.

[The Sketch Engine] The Sketch Engine project: http://www.sketchengine.co.uk/

[Rychly, Smrz, 2004] Rychly, P., Smrz, P. Manatee, Bonito and Word Sketches for Czech. (2004). In *Trudy mezhdunarodnoy konferentsii "Korpusnaya lingvistika-2004": Sbornik dokladov*. St.-Petersburg, 324-334.

[Mel'cuk, 1998] Mel'cuk, I.A. (1998). Collocations and Lexical Functions. In Cowie, A.P. (ed.): *Phraseology: Theory, Analysis, and Applications*. Oxford. P. 23-53.

[Bol'shakova et al., 2007] Bol'shakova E.I., Baeva N.V., Bordachenkova E.A., Vasil'eva N.Je., Morozov S.S. Lek-siko-sintaksicheskie shablony v zadachah avtomaticheskoj obrabotki tekstov. In *Kom-p'juternaya lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferencii «Dialog 2007»* (Bekasovo, 30 maya - 3 iyunya 2007 g.). Vyp. 6 (13). M.: RGGU, 2007. S. 70-75.

[Mitrofanova, Zakharov, 2009] Mitrofanova O.A., Zakharov V.P. Avtomatizirovannyj analiz terminologii v russkoyazychnom korpuse tekstov po korpusnoj lingvistike. In *Komp'juternaya lingvistika i intellektual'nye tehnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferencii «Dialog 2009»* (Bekasovo, 27-31 maya 2009 g.). Vyp. 8 (15). M.: RGGU, 2009. S. 321-328.

[Russkaya grammatika, 1980] Russkaya grammatika. Ju. Shvedova (ed.). T. I, II. M.: Nauka, 1980.

[Zolotova, 1988] Zolotova G.A. Sintaksicheskij slovar': Repertuar jelementarnyh edinic russkogo sintaksisa. M.: Nauka, 1988.

[Bol'shakov, Bol'shakova, 2006] Bol'shakov I.A., Bol'shakova E.I. Rasshirennyj eksperiment po avtomaticheskomu obnaruzheniju i ispravleniju russkih malapropizmov. In *Komp'juternaya lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferencii «Dialog 2006»* (Bekasovo, 31 maya – 4 iyunya 2006 g.). M., 2006. S. 78–83.

Authors' Information

***Maria Khokhlova** – Assistant Professor, Saint-Petersburg State University, Universitetskaya emb., 11, Saint-Petersburg 199034 Russia; e-mail: khokhlova.marie@gmail.com*

Major Fields of Scientific Research: Natural Language Processing

***Victor Zakharov** – Associate Professor, Saint-Petersburg State University, Universitetskaya emb., 11, Saint-Petersburg 199034 Russia; e-mail: vz1311@yandex.ru*

Major Fields of Scientific Research: Corpus linguistics, Computational Lexicography