Automatic Detection of Stable Grammatical Features in N-Grams

Mikhail Kopotev[1] **Lidia Pivovarova**[1,2] **Natalia Kochetkova**[3] **Roman Yangarber**[1]
[1] University of Helsinki, Finland
[2] St.Petersburg State University, Russia
[3] Moscow Institute of Electronics and Mathematics, NRU HSE, Russia

Abstract

This paper presents an algorithm that allows the user to issue a query pattern, collects multi-word expressions (MWEs) that match the pattern, and then ranks them in a uniform fashion. This is achieved by quantifying the strength of all possible relations between the tokens and their features in the MWEs. The algorithm collects the frequency of morphological categories of the given pattern on a unified scale in order to choose the stable categories and their values. For every part of speech, and for all of its categories, we calculate a normalized Kullback-Leibler divergence between the category's distribution in the pattern and its distribution in the corpus overall. Categories with the largest divergence are considered to be the most significant. The particular values of the categories are sorted according to a frequency ratio. As a result, we obtain morpho-syntactic profiles of a given pattern, which includes the most stable category of the pattern, and their values.

1 Introduction

In n-grams, the relations among words and among their grammatical categories cover a wide spectrum, ranging from idioms to syntactic units, such as a verb phrase. In most cases, the words are linked together by both grammatical and lexical relations. It is difficult to decide, which relation is stronger in each particular case. For example, in the idiomatic phrase *meet the eye*, the relationship is lexical rather than grammatical. A phrasal verb *meet up* is similar to single-word verbs and has its own meaning. It can be interpreted as one lexeme, spelled as two words.

On the other hand, phrases like *meet the requirements*, *meet the specifications*, *meet the demands* are traditionally called "collocations." However, the question arises about the role played by the noun following the verb: is it a lexically free direct object, or a part of stable lexical unit, or to some extend both? These words are bound by both grammatical and lexical relations, and we assume that the majority of word combinations in any language have such a dual nature.

Lastly, the relationship between the words in the English phrase *meet her* differs from those above in that it may be described as purely grammatical—the verb *meet* receives a direct object.

Distinguishing *collocations*, i.e. "co-occurrences of words" from *colligations*, i.e. "co-occurrence of word forms with grammatical phenomena" (Gries and Divjak, 2009) is not always a simple task; there is no clear boundary between various types of word combinations inasmuch as they can be simultaneously a collocation and a colligation—this type of MWE is called *collostructions* in (Stefanowitsch and Gries, 2003). It was proposed that language as such is a "constructicon" (Goldberg, 2006), which means that fusion is its core nature. For this reason, devising formal methods to measure the strength of morphological or lexical relations between words becomes a challenge.

Our approach aims to treat multi-word expressions (MWEs) of various nature—idioms, multi-word lexemes, collocations and colligations—*on an equal basis*, and to compare the strength of various possible relations between the tokens in a MWE quantitatively. We search for "the underlying cause"

Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013), pages 73–81,
Atlanta, Georgia, 13-14 June 2013.

for the frequent co-occurrence of certain words: whether it is due to their morphological categories, or lexical compatibility, or a combination of both. In this paper, however, we focus on colligations, ignoring collocations and collostructions.

For languages with rich morphology the situation is more complicated, because each word may have several morphological categories that are not independent and interact with each other. This paper focuses on Russian, which not only has free word order and rich morphology,[1] but is also a language that is well-investigated. A good number of corpora and reference grammars are available to be used for evaluation. The data we use in this work is the n-gram corpus, extracted from a deeply annotated and carefully disambiguated (partly manually) sub-corpus of the Russian National Corpus (RNC). The size of disambiguated corpus used in this paper is 5 944 188 words of running text.

2 Related Work

Much effort has been invested in automatic extraction of MWEs from text. A great variety of method are used, depending on the data, the particular tasks and the types of MWEs to be extracted. Pecina (2005) surveys 87 statistical measures and methods, and even that is not a complete list. The most frequently used metrics, *inter alia*, are Mutual Information (MI), (Church and Hanks, 1990), t-score (Church et al., 1991), and log-likelihood (Dunning, 1993). The common disadvantage of these is their dependency on the number of words included in the MWE. Although there is a large number of papers that use MI for bigram extraction, only a few use the MI measure for three or more collocates, e.g., (Tadić and Šojat, 2003; Wermter and Hahn, 2006; Kilgarriff et al., 2012),

Frantzi et al. (2000) introduced the c-value and nc-value measures to extract terms of different lengths. Daudaravicius (2010) has developed a promising method that recognizes collocations in text. Rather than extracting MWEs, this method cuts the text into a sequence of MWEs of length from 1 to 7 words; the algorithm may produce different chunking for the same segment of text within different corpora. Nevertheless, extraction of variable-length MWE is a challenging task; the majority of papers in the field still use measures that take the number of collocates as a core parameter.

Entropy and other probabilistic measures have been used for MWE extraction since the earliest work. For example, the main idea in (Shimohata et al., 1997; Resnik, 1997), is that the MWE's idiosyncrasy, (Sag et al., 2002), is reflected in the distributions of the collocates. Ramisch et al. (2008) introduced the Entropy of Permutation and Insertion:

$$EPI = -\sum_{a=0}^{m} p(ngram_a) \log[p(ngram_a)] \quad (1)$$

where $ngram_0$ is the original MWE, and $ngram_a$ are its syntactically acceptable permutations. Kullback-Leibler divergence was proposed by Resnik (1997) to measure selective preference for the word sense disambiguation (WSD) task. Fazly and Stevenson (2007) applied a set of statistical measures to classify *verb+noun* MWEs and used Kullback-Leibler divergence, among other methods, to measure the syntactic cohesion of a word combination. Van de Cruys and Moirón (2007) used normalized Kullback-Leibler divergence to find idiomatic expression with verbs in Dutch.

Russian MWE-studies have emerged over the last decade. Khokhlova and Zakharov (2009) applied MI, t-score and log-likelihood to extract verb collocations; Yagunova and Pivovarova (2010) studied the difference between Russian lemma/token collocations and also between various genres; Dobrov and Loukachevitch (2011) implemented term extraction algorithms. However, there is a lack of study of both colligations and collostructions in Russian. The only work known to us is by Sharoff (2004), who applied the MI-score to extract prepositional phrases; however, the only category he used was the POS.

As far as we aware, the algorithm we present in this paper has not been applied to Russian or to other languages.

3 Method

The input for our system is any n-gram of length 2–4, where one position is a gap—the algorithm aims

[1] The Multitext-East specification, which aims to create an unified cross-language annotation scheme, defines 156 morpho-syntactic tags for Russian as compared to 80 tags for English (http://nl.ijs.si/ME/V4/msd/html).

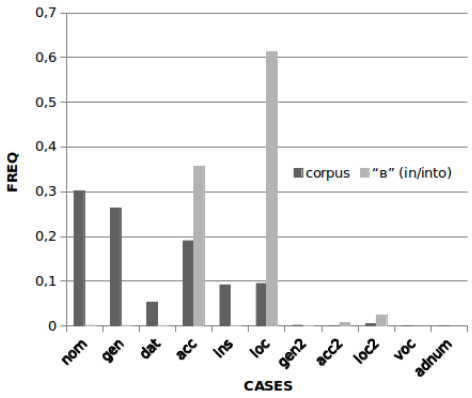

Figure 1: Distributions of noun cases in the corpus and in a sample—following the preposition "в" (*in*)

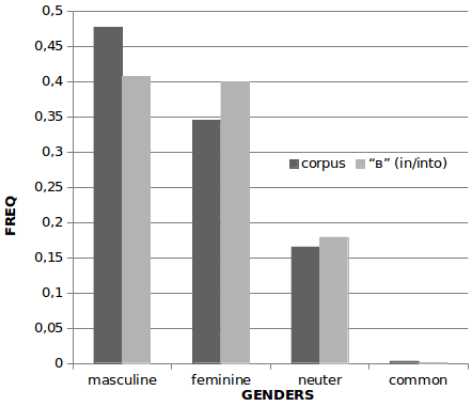

Figure 2: Distributions of nominal gender in the corpus and in a sample—following the preposition "в" (*in*)

to find the most stable morphological categories of words that can fill this gap. Moreover, the user can specify the particular properties of words that can fill the gap—for example, specify that the output should include only plural nouns. Thus, the combination of the surrounding words and morphological constrains form an initial query *pattern* for the algorithm.

Our model tries to capture the difference between distributions of linguistic features in the general corpus as compared to distributions within the given pattern. For example, Figure 1 shows the distribution of cases in the corpus overall vs. their distribution in words following the preposition "в" (*in/into*). Figure 2 shows the corresponding distributions of gender. Gender is distributed similarly in the corpus and in the sample restricted by the pattern; by contrast, the distribution of cases is clearly different. This is due to the fact that the preposition governs the case of the noun, but has no effect on gender. To measure this difference between the distributions we use the Kullback-Leibler divergence:

$$Div(C) = \sum_{i=1}^{N} P_i^{pattern} \times \log(\frac{P_i^{pattern}}{P_i^{corpus}}) \qquad (2)$$

where C is the morphological category in a pattern—e.g., case or gender,—having the values $1..N$, $P_i^{pattern}$ is the relative frequency of value i restricted by the pattern, and P_i^{corpus} is the relative frequency of the same value in the general corpus. Since the number of possible values for a category is variable—e.g., eleven for case, four for gender, and hundreds of thousands for lemmas—the divergence needs to be normalized. The normalization could be done in various ways, e.g., against the entropy or some maximal divergence in the data; in our experiments, the best results were obtained using a variant proposed in (Bigi, 2003), where the divergence between the corpus distribution and the uniform distribution is used as the normalizing factor:

$$NormDiv(C) = \frac{Div(C)}{E(C) + \log(n)} \qquad (3)$$

where $E(C)$ is the entropy of category C and n is the number of possible values of C; the term $\log(n)$ is the entropy of the uniform distribution over n outcomes (which is the maximal entropy). The category with the highest value of normalized divergence is seen as maximally preferred by the pattern.

However, divergence is unable to determine the exact *values* of the category, and some of these values are clearly unreliable even if they seem to appear in the pattern. For example, Figure 1 shows that preposition "в" (*in*) in the data is sometimes followed by the nominative case, which is grammatically impossible. This is due to a certain amount of noise, which is unavoidable in a large corpus due to mark-up errors or inherent morphological ambiguity. In Russian, the nominative and accusative cases often syncretize (assume identical forms), which can cause inaccuracies in annotation. On the other hand, some values of a category can be extremely rare; thus, they will be rare within patterns as well. For instance, the so-called "second accusative" case (labeled "acc2" in Figure 1) is rare in modern Russian,

which is why its appearance in combination with preposition "в" (*in*) is significant, even though its frequency is not much higher than the frequency of the (erroneous) nominative case in the same pattern.

To find the significant values of a particular category we use the ratio between the frequencies of the value in a sample and in the corpus:

$$frequency\_ratio = \frac{P_i^{pattern}}{P_i^{corpus}} \quad (4)$$

If $frequentcy\_ratio > 1$, then the category's value is assumed to be *selected* by the pattern.

Finally, we note that the distribution of POS varies considerably within every pattern as compared to its distribution in the corpus. For example, prepositions can be followed only by noun groups and can never be followed by verbs or conjunctions. This means the Kullback-Leibler divergence for any POS, naturally assumes the highest value in *any* pattern; for this reason, we exclude the POS category from consideration in our calculation, aiming to find more subtle and interesting regularities in the data.

To summarize, the algorithm works as follows: for a given *query pattern*

1. search all words that appear in the query pattern and group them according to their POS tags.

2. for every POS, calculate the normalized Kullback-Leibler divergence for all of its categories; categories that show the maximum divergence are considered to be the most significant for the given pattern;

3. for every relevant category, sort its values according to the frequency ratio; if frequency ratio is less than 1, the value considered to be irrelevant for this pattern.

4 Experiments

In this paper, we conduct an in-depth evaluation focusing on a limited number of linguistic phenomena, namely: bigrams beginning with single-token prepositions, which impose strong morpho-syntactic constraints in terms of case government. We investigate 25 prepositions, such as "без" (*without*), "в" (*in/to*), etc. We evaluate the corpus of bigrams systematically against these queries, although we expect that the model we propose here produces relevant results for a much wider range of constructions—to be confirmed in further work.

4.1 Prepositions and Morphological Category

A syntactic property of prepositions in Russian is that they govern nominal phrases, i.e., that we expect the largest normalized divergence in queries such as { Preposition + X }, where the POS of X is *noun*, to occur exactly with the category of case. Figure 3 shows the normalized divergence for four lexical and morphological categories. Among them, Case has the maximal divergence for all prepositions, which matches our expectation with 100% accuracy.

According to the figure, the morphological category of Animacy[2] is also interesting, in that it has a high value for some prepositions, like "из-под" (*from under*), "под" (*under*), "над" (*above*). A good example is the preposition "из-под" (*from under*). Its semantic properties cause inanimate nouns to appear much more frequently than animate ones. Consequently, we observe a higher divergence, due to inanimate nouns like "из-под земли" (*from under ground*), "из-под снега" (*from under the snow*), etc. Another good example of hidden semantic properties is a pair of prepositions "под" (*under*) and "над" (*above*). One can expect that their syntactic behaviour is more or less similar, but the histogram shows that Animacy (surprisingly) has a much higher divergence for "под" (*under*) to be ignored. Indeed, a deeper corpus-based analysis reveals a stable, frequently used construction, which gives many points to animate nouns, e.g., "замаскированный под невесту" (*disguised as a bride*). It is notable that this particular effect is not mentioned in any grammar book, (to the best of our knowledge).

To conclude, the Case category is the clear winner in terms of having the greatest normalized divergence, and the output fully matches the expectation on all 25 common prepositions that we tested. Other results are also clearly interesting due to their links to semantic properties, that is, to collocations. The next task is, therefore to discriminate

[2] Animacy is a morphological category of Russian nouns based on whether the referent of the noun is considered sentient or living. Most nouns denoting humans and animals are animate, while the majority of other nouns are inanimate.

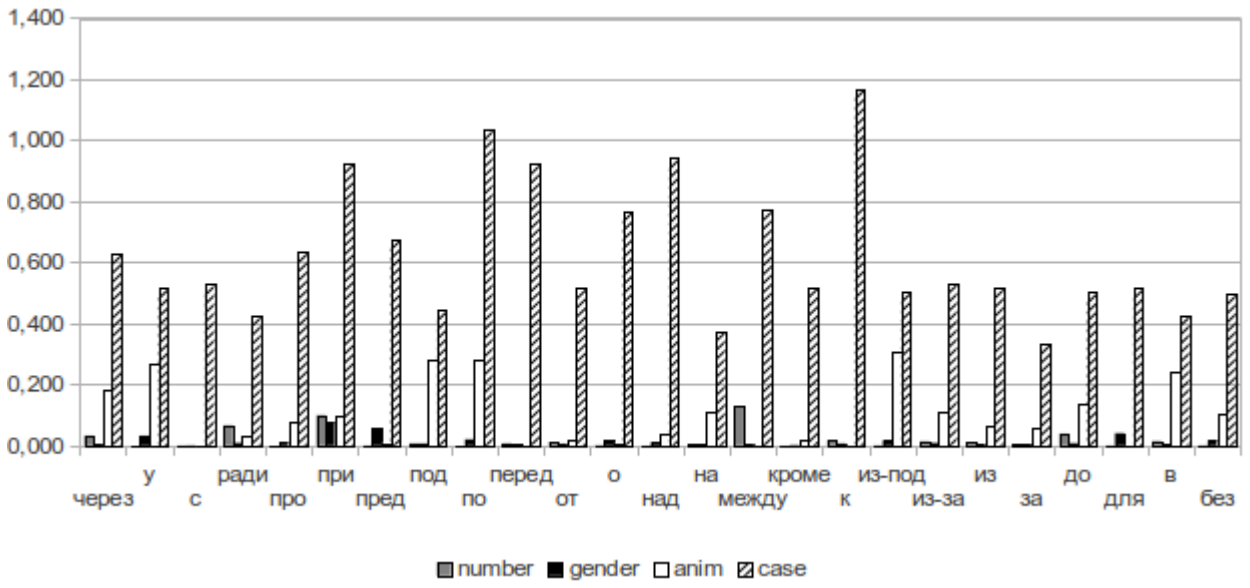

Figure 3: Normalized divergence of noun categories (grammemes) for pattern *preposition+X*.

between the runners-up, like Animacy for "под" (*under*), which seem to be interesting to some extent, and clear losers like Gender, in the example above. To do that we need to find an appropriate threshold—preferably automatically—between relevant and non-relevant results. The algorithm ranks the categories according to their divergence; the category that has the top rank is certainly meaningful. The question is how to determine which among the rest are significant as well; this is left for future work.

4.2 Specific Values of the Category with Maximum Divergence

The next question we explore is which particular *values* of the maximally divergent category—here, Case—are selected by a given preposition. As we mentioned above, we use the frequency ratio for this task. We collected a list of cases[3] that appear after the given preposition, according to the algorithm with $frequency\_ratio > 1$; which cases are possible according to grammatical descriptions,[4] which cases were produced by the algorithm, and the number of correct cases in the system's response. We expect that by using the frequency ratio we can reduce the noise; for example, of the eight cases that match the pattern { "с" (*with*) + Noun } only four are relevant.

The algorithm predicts the correct relevant set for 21 of 25 prepositions, giving a total precision of 95%, recall of 89%, and F-measure of 92%. The prepositions highlighted in bold in Table 1 are those that were incorrectly processed for various reasons; the error analysis is presented below.

14: "о" (*about*) The algorithm unexpectedly flags the *voc* (vocative) as a possible case after this preposition. This is incorrect; checking the data we discovered that this mistake was due to erroneous annotation: the interjection "о" (*oh*), as in "О боже!" (*Oh God!*), is incorrectly annotated as the preposition "о" (*about*). The error occurs twice in the data. However, as the vocative is extremely rare in the data (its frequency in the corpus is less than 0,0004), two erroneous tags are sufficient to give it a high rank. Similar annotation errors for more frequent cases are eliminated by the algorithm. For example, as we mentioned in the previous section, the nominative

[3]The current annotation scheme of our data has eleven case tags, namely: nom, voc, gen, gen2, dat, acc, acc2, ins, loc, loc2, adnum.

[4]Note, that not all possible prep+case combinations are represented in the corpus; for example, the combination { "ради" (*for the sake of*) + gen2 } does not appear in our data, and only eight times in the RNC overall. For evaluation we take into consideration only those prep+case combinations that appear at least once in our dataset.

	Preposition	Meaning	Expected cases	Response
1	без	*without*	gen/gen2	gen/gen2
2	в	*in/into*	acc/acc2/loc/loc2	acc/acc2/loc/loc2
3	для	*for*	gen/gen2	gen/gen2
4	до	*until*	gen/gen2	gen/gen2
5	за	*behind*	acc/ins	acc/ins
6	из	*from*	gen/gen2	gen/gen2
7	из-за	*from behind*	gen/gen2	gen/gen2
8	из-под	*from under*	gen/gen2	gen/gen2
9	к	*to*	dat	dat
10	кроме	*beyond*	gen	gen
11	между	*between*	ins	ins
12	на	*on*	acc/loc/loc2	acc/loc/loc2
13	над	*above*	ins	ins
14	о	*about*	**acc/loc**	**loc/voc**
15	от	*from*	gen/gen2	gen/gen2
16	перед	*in front of*	ins	ins
17	пред	*in front of*	ins	ins
18	по	*by/up to*	**dat/loc/acc**	**dat**
19	под	*under*	acc/ins	acc/ins
20	при	*at/by*	loc	loc
21	про	*about*	acc	acc
22	ради	*for*	gen	gen
23	с	*with*	**gen/gen2/acc/ins**	**gen2/ins**
24	у	*near*	gen	gen
25	через	*through*	**acc**	**acc/adnum**

Expected	45	**Precision**	0.95
Response	42	**Recall**	0.89
Correct	40	**F-measure**	0.92

Table 1: Noun cases expected and returned by the algorithm for Russian prepositions.

case after preposition "в" (*in*) appears 88 times in our data; however this case is not returned by the algorithm, since it is below the frequency ratio threshold.

25: "через" (*through/past*) The adnumerative (adnum) is a rare case in our data, so even a single occurrence in a sample is considered important by the algorithm. A single bigram is found in the data, where the token "часа" (*hours*)—correctly annotated with the *adnum* tag—predictably depends on the Numeral, i.e., "два" (*two*), rather than on preposition "через" (*through/past*), see Figure 4. The numeral appears in *post-position*—a highly marked word order that is admissible in this colloquial construction in Russian: "через часа два" (*lit.: after hours two = idiom: after about two hours*), where

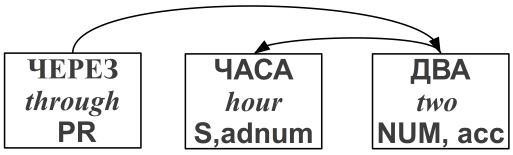

Figure 4: Distributions of cases in the corpus and in a sample. (Arrows indicate syntactic dependency.)

the preposition governs the Case of the numeral, and the numeral governs a noun that *precedes* it.

Because our algorithm at the moment processes linear sequences, these kinds of syntactic inversion phenomena in Russian will pose a challenge. In general this problem can be solved by using tree-banks for MWE extraction, (Seretan, 2008; Martens and Vandeghinste, 2010). However, an appropriate tree-

bank is not always available for a given language; in fact, we do not have access to any Russian tree-bank suitable for this task.

23: "с" (*with*) This is a genuine error. The algorithm misses two of four correct cases, Genitive and Accusative, because both are widely used across the corpus, which reduces their frequency ratio in the sub-sample. Our further work will focus on finding flexible frequency ratio thresholds, which is now set to one. Two of the correct cases (Instrumental and Gen2) are well over the threshold, while Genitive, with 0.6924, and Accusative, with 0.0440, fall short.

18: "по" (*by/along*) For this preposition the algorithm predicts 1 case out of 3. This situation is slightly different from the previous ones, since the accusative and locative cases are much more rare with preposition "по" (*by/along*) than the dative: 245 instances out of 15387 for accusative, and 222 for locative in our data. We hypothesize that this means that such "Prep+case" combinations are constrained lexically to a greater extent than grammatically. To check this hypothesis we calculate the frequency ratio for all lemmas that appear with the respective patterns { "по" (*by/along*) + acc } and { "по" (*by/along*) + loc }. As a result, 15 distinct lemmas were extracted by { "по" (*by*) + acc }; 13 out of them have $frequency\_ratio > 1$. The majority of the lemmas belong to the semantic class "part of the body" and are used in a very specific Russian construction, which indicates "an approximate level", e.g. "по локоть" (*up to (one's) elbow*), cf. English *"up to one's neck in work"*. This construction has limited productivity, and we are satisfied that the Accusative is omitted in the output for grammatical categories, since the algorithm outputs all tokens that appear in the { "по" (*by/along*) + acc } as relevant lemmas.

The case of { "по" (*by*) + loc } is more complex: 44 of 76 combinations return a frequency greater than 1. Analysis of annotation errors reveals a compact collection of bureaucratic cliches, like "по прибытии" (*upon arrival*), "по истечении" (*upon completion*), etc., which all share the semantics of "*immediately following X*", and are pragmatically related. These are expressions belonging to the same bureaucratic jargon and sharing the same morphological pattern, however, they are below the threshold. Again, we are faced with need to tune the threshold to capture this kind of potentially interesting lexical combinations. In general, semantic and pragmatic factors influence the ability of words to combine, and the algorithm shows it in some way, though these aspects of the problem are beyond the scope of our experiments in the current stage.

5 Discussion and Future Work

5.1 Development of the algorithm

We have presented a part an overall system under development. In the preceding sections, we investigate an area where collocations and colligations meet. To summarize, the algorithm, based on the corpus of n-grams, treats both morpho-syntactic and lexical co-occurrences as a unified continuum, which has no clear borders. The evaluation of the morphological output raises some new questions for further development:

- At present, the low precision for both low- and high-frequency tags depends on the threshold, which needs to be studied further.
- The values of divergences are currently not normalized among the different query patterns. This may be a difficult question, and we plan to investigate this further. The algorithm provides a way to compare the strength of very diverse collocations, which have nothing in common, in terms of their degree of idiomatization.
- We observe that the longer the n-gram, the more we expect it to be a collocation; stable bigrams appear more frequently to be colligations, while stable 4-grams are more often collocations. The problem is that those collocations with a highly frequent first collocate, e.g., "в" (*in*), cannot be found using our algorithm as it stands now.
- Token/lexeme stability is the next task we will concentrate on. Wermter and Hahn (2006) and Kilgarriff et al. (2012) proposed that sorting tokens/lexemes according to plain frequency works well if there is no grammatical knowledge at hand. We do have such knowledge. To improve the accuracy of lexeme/token extraction we rely on the idea of grammatical pro-

files, introduced by Gries and Divjak (2009). We plan to develop this approach with the further assumption that the distribution of tokens/lexemes within a pattern is based on relevant grammatical properties, which are obtained in an earlier step of our algorithm. For instance, for "не до X" (*not up to X*) we have found that the grammatical profile for X is N.gen/gen2, and the token $frequency\_ratio$ is greater than 1 as well. Building the list of tokens that are the most stable for this pattern, we compare their distributions within the pattern to all N.gen/gen2 tokens in the corpus. This yields the following tokens as the most relevant: "не до смеха" (*lit.: not up to laughter.gen = idiom: no laughing matter*);"не до жиру" (*lit. not up to fat.gen2 = idiom: no time/place for complacency*), which reveals an interesting set of idioms.

5.2 Extensions and Applications

The model has no restriction on the length of data to be used, and is applicable to various languages. Finnish (which is morphologically rich) and English (morphologically poor) will be examined next. As for Russian, so far the algorithm has been systematically evaluated against bigrams, although we have 3-, 4- and 5-grams at our disposal for future work.

A reliable method that is able to determine patterns of frequently co-occurring lexical and grammatical features within a corpus can have far-reaching practical implications. One particular application that we are exploring is the fine-tuning of semantic patterns that are commonly used in information extraction (IE), (Grishman, 2003). Our work on IE focuses on different domains and different languages, (Yangarber et al., 2007; Atkinson et al., 2011). Analysis of MWEs that occur in extraction patterns would provide valuable insights into how the patterns depend on the particular style or *genre* of the corpus, (Huttunen et al., 2002). Subtle, genre-specific differences in expression can indicate whether a given piece of text is signaling the presence an event of interest.

5.3 Creating Teaching-Support Tools

Instructors teaching a foreign language are regularly asked how words co-occur: What cases and word forms appear after a given preposition? Which ones should I learn by rote and which ones follow rules? The persistence of such questions indicates that this is an important challenge to be addressed—we should aim to build a system that can automatically generate an integrated answer. A tool that produces answers to these questions would be of great help for teachers as well as students. The presented algorithm can support an easy-to-use Web-based application, or an application for a mobile device. We plan to develop a service, which is able to process queries described in the paper. This service would be an additional interface to a corpus, aimed at finding not only the linear context of words but also their collocational and constructional preferences. We believe that such an interface would be useful for both research and language-learning needs.

Acknowledgments

We are very grateful to the Russian National Corpus developers, especially E. Rakhilina and O. Lyashevskaya, for providing us with the data.

References

Martin Atkinson, Jakub Piskorski, Erik van der Goot, and Roman Yangarber. 2011. Multilingual real-time event extraction for border security intelligence gathering. In U. Kock Wiil, editor, *Counterterrorism and Open Source Intelligence*, pages 355–390. Springer Lecture Notes in Social Networks, Vol. 2, 1st edition.

Brigitte Bigi. 2003. Using Kullback-Leibler distance for text categorization. In Fabrizio Sebastiani, editor, *Advances in Information Retrieval*, volume 2633 of *Lecture Notes in Computer Science*, pages 305–319. Springer Berlin, Heidelberg.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Kenneth Church, William Gale, Patrick Hanks, and Donald Kindle. 1991. Using statistics in lexical analysis. *Lexical acquisition: exploiting on-line resources to build a lexicon*.

Vidas Daudaravicius. 2010. Automatic identification of lexical units. *Computational Linguistics and Intelligent text processing CICling-2009*.

Boris Dobrov and Natalia Loukachevitch. 2011. Multiple evidence for term extraction in broad domains. In *Proceedings of the 8th Recent Advances in Natural Language Processing Conference (RANLP 2011). Hissar, Bulgaria*, pages 710–715.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.

Afsaneh Fazly and Suzanne Stevenson. 2007. Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 9–16. Association for Computational Linguistics.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Adele Goldberg. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press, USA.

Stefan Th. Gries and Dagmar Divjak. 2009. Behavioral profiles: a corpus-based approach to cognitive semantic analysis. *New directions in cognitive linguistics*, pages 57–75.

Ralph Grishman. 2003. Information extraction. In *The Handbook of Computational Linguistics and Natural Language Processing*, pages 515–530. Wiley-Blackwell.

Silja Huttunen, Roman Yangarber, and Ralph Grishman. 2002. Diversity of scenarios in information extraction. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas de Gran Canaria, Spain, May.

Maria Khokhlova and Viktor Zakharov. 2009. Statistical collocability of Russian verbs. *After Half a Century of Slavonic Natural Language Processing*, pages 125–132.

Adam Kilgarriff, Pavel Rychlỳ, Vojtech Kovár, and Vıt Baisa. 2012. Finding multiwords of more than two words. In *Proceedings of EURALEX2012*.

Scott Martens and Vincent Vandeghinste. 2010. An efficient, generic approach to extracting multi-word expressions from dependency trees. In *CoLing Workshop: Multiword Expressions: From Theory to Applications (MWE 2010)*.

Pavel Pecina. 2005. An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL Student Research Workshop*, pages 13–18. Association for Computational Linguistics.

Carlos Ramisch, Paulo Schreiner, Marco Idiart, and Aline Villavicencio. 2008. An evaluation of methods for the extraction of multiword expressions. In *Proceedings of the LREC Workshop-Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 50–53.

Philip Resnik. 1997. Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*, pages 52–57. Washington, DC.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. *Computational Linguistics and Intelligent Text Processing*, pages 189–206.

Violeta Seretan. 2008. *Collocation extraction based on syntactic parsing*. Ph.D. thesis, University of Geneva.

Serge Sharoff. 2004. What is at stake: a case study of Russian expressions starting with a preposition. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 17–23. Association for Computational Linguistics.

Sayori Shimohata, Toshiyuki Sugio, and Junji Nagata. 1997. Retrieving collocations by co-occurrences and word order constraints. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 476–481. Association for Computational Linguistics.

Anatol Stefanowitsch and Stefan Th Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics*, 8(2):209–243.

Marko Tadić and Krešimir Šojat. 2003. Finding multiword term candidates in Croatian. In *Proceedings of IESL2003 Workshop*, pages 102–107.

Tim Van de Cruys and Begona Villada Moirón. 2007. Lexico-semantic multiword expression extraction. In *Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands (CLIN)*, pages 175–190.

Joachim Wermter and Udo Hahn. 2006. You can't beat frequency (unless you use linguistic knowledge) – a qualitative evaluation of association measures for collocation and term extraction. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 785–792.

Elena Yagunova and Lidia Pivovarova. 2010. The nature of collocations in the Russian language. The experience of automatic extraction and classification of the material of news texts. *Automatic Documentation and Mathematical Linguistics*, 44(3):164–175.

Roman Yangarber, Clive Best, Peter von Etter, Flavio Fuart, David Horby, and Ralf Steinberger. 2007. Combining information about epidemic threats from multiple sources. In *Proceedings of the MMIES Workshop, International Conference on Recent Advances in Natural Language Processing (RANLP 2007)*, Borovets, Bulgaria, September.