Mémoires de la Société Néophilologique de Helsinki

Tome C

Collocations Cross-Linguistically: Corpora, Dictionaries and Language Teaching

ed. by
Begoña Sanromán Vilas

Helsinki 2016
Société Néophilologique

ISBN-10: 951-9040-57-9
ISBN-13: 978-951-9040-57-8
EAN: 9789519040578
Helsinki 2016

Contents

Constructional generalization over Russian collocations

Mikhail Kopotev, University of Helsinki
Daria Kormacheva, University of Helsinki
Lidia Pivovarova, University of Helsinki

The CoCoCo project aims to model multi-word expressions (MWEs) of diverse natures in a unified fashion. The algorithm predicts the most stable features in an n-gram—morphological, lexical, or constructional. In this article, we focus more on lexical compatibility of extracted collocations. At one extreme are lexically stable idioms, where no generalization is possible, e.g., *lo and behold*. Other collocations appear to be stable on a more abstract level of generalization. They are constructions where lexical items are replaceable but belong to the same semantic class, e.g., *sleight of [hand/mouth/mind]*. In this case, prediction of the entire semantic class is possible. To confirm this idea, we present a qualitative analysis of automatically extracted Russian MWEs. We then use distributional semantics methods to find semantic classes automatically and demonstrate that these correspond with manually annotated classes. This implies that the semantic classes can be used in the collocation detection algorithm.

1. Introduction

A speech act is produced linearly: after saying A, we may be more likely to say B rather than C. In the flow of speech, many word combinations can be identified as not idiomatic in a narrow sense, but rather as sharing a common property—a stable co-occurrence. These repeatedly co-occurring items potentially develop into idioms, single-word tokens or even morphemes. However, prior to being coined into more fixed items, co-occurrences exist as a set of ready-to-use prefabricated chunks (Hunston & Francis 2000; Sinclair & Mauranen 2006). These broad and rather poorly defined word combinations are often called *multiword expressions* (MWEs). In our project CoCoCo, we take into account more strict types of such MWEs: lexical associations (**collocations**[1]), as well as grammatical and seman-

1 "Collocation typically denotes frequently repeated or statistically significant co-occurrences, whether or not there are special semantic bonds between collocating items" (Moon 1998: 26).

tic associations (**colligations**[2] and **constructions**,[3] respectively). The full range of such associations includes grammatical and lexical features—without drawing *ad hoc* borders between these—that can be determined by context. Among these kinds of associations, collocations play a fundamental role, because they are what we actually have in a corpus—a string of words.

> The MWE *bez galstuka* (lit. 'without a neck-tie') can be described at multiple levels. First, it exemplifies the grammatically restricted colligation [*bez* 'without'+ Noun.GEN]; secondly it represents the semantic preferences of a construction [*bez* 'without' + Noun.GEN 'clothing item']. Finally, this is an idiom meaning 'informal' (like *vstreča bez galstuka* 'an informal/shirtsleeve meeting').

In our project we aim to model MWEs of various natures on an equal basis. Our algorithm compares the strength of various possible relations between the tokens in a given n-gram, a linear sequence of tokens, and searches for the underlying cause that binds the words together, whether this be lexical, grammatical, or a combination of both. Taking syntactic, semantic, and lexical properties equally into account, we follow the ideas that were first formulated by J. Sinclair, A. Goldberg, and Ch. Fillmore et al. (1988) and developed recently by S. Gries & A. Stefanowitsch (2004) and Hunston & Francis (2000), to mention just a few. However, in this article, we focus more on lexical compatibility of extracted collocations. We investigate two reasons why lexical items co-occur, namely, the tendency of certain collocations to be idiomatic and the proneness of collocates to cluster as having overlapping semantic preferences within a certain construction; in this paper, we focus mostly on constructions.

2. Overview of the Algorithm

Rather than applying a single multiword-extraction technique, we propose a cascade of procedures that builds on the results of the preceding steps. The general overview of the algorithm is shown in Fig. 1 (the part that is the focus of this article is highlighted in gray).

2 "The grammatical company a word keeps (or avoids keeping) and the positions it prefers" (Hoey 2004: 28).

3 "A pairing of form with meaning/use such that some aspect of the form or some aspect of the meaning/use is not strictly predictable" (Goldberg 1996: 68).

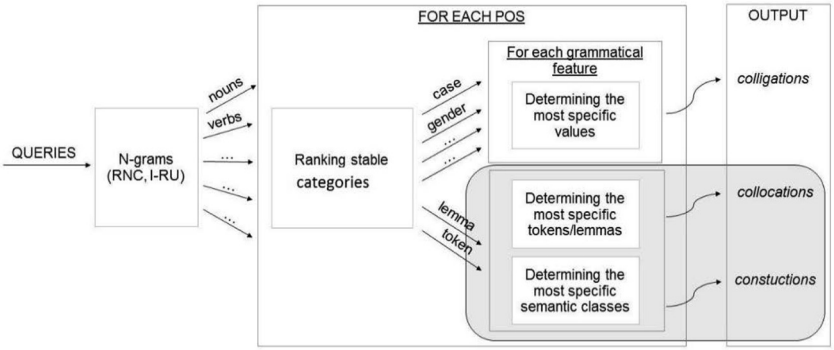

Figure 1. General overview of the algorithm (POS = parts of speech)

The algorithm works as follows: The system takes as input a query—an n-gram (currently of length 2-4) with possible grammatical constraints, where one of the positions is a sought variable. Thus, the query is a pattern. The aim is to find the most stable lexical and grammatical features of the values that can appear in this pattern. The data we use consist of n-grams extracted from a carefully disambiguated subcorpus of the Russian National Corpus (RNC, Rakhilina 2005); for comparison (in Part 4), we also use the Russian Internet Corpus by S. Sharoff (I-RU, Sharoff & Nivre 2011).

First, the algorithm finds all words in the corpus that match the pattern and groups them according to their part-of-speech (POS) tags. Then, for every POS, the system determines the most stable *features*, which include grammatical categories (case, gender, etc.), tokens, and lemmas. To find the most stable features, we exploit the difference between the distribution of the *feature values* in the pattern vs. the distribution in the corpus as a whole. This difference is measured using Kullback-Leibler divergence (KLD):

$$D_{KL}(Q_C \parallel P_C) = \sum_{i=1}^{N} Q(c_i) \times log\left(\frac{Q(c_i)}{P(c_i)}\right)$$

where C is a category in the pattern, having the values $1..N$, $Q(c_i)$ is the relative frequency of value i restricted by the pattern, and $P(c_i)$ is the relative frequency of the same value in the corpus overall.

Since the number of possible values is different for the different categories/features, we use normalized Kullback-Leibler divergence:

$$NormD_{KL}(Q_C \parallel P_C) = \frac{D_{KL}(Q_C \parallel P_C)}{H(P_C) + log(N)}$$

where $H(P_C)$ is the entropy of category *C*; *log(N)* is the entropy of the uniform distribution over *N* outcomes (which is the maximum entropy). The feature with the highest value of normalized divergence is seen as maximally preferred by the pattern. For a more detailed explanation and evaluation of this measure, please refer to Kopotev et al. (2013).

Using normalized KLD, we obtain a ranked list of grammatical categories, tokens, and lemmas. For example, the query pattern [*bez* 'without' + *Noun*] results in the following ranking of categories/features for the *Noun* variable: 1) Case: 0.3195; 2) Token: 0.1608; 3) Lemma: 0.1123; 4) <…>; 5) Gender: 0.0159; 6) Number: 0.0008.

The most stable category for *Noun* in this pattern is *Case*, which matches the linguistic intuition well, because prepositions govern the case of the dependent NP. At the same time, the last two categories—gender and number—are about equally distributed in the corpus overall and in the instances matched by the query, and thus do not depend on the preposition.

Having specified the most stable categories, we define particular *values* for them. In this step, grammatical categories are processed separately from tokens and lemmas, since tokens and lemmas have significantly different distributional properties than grammatical categories; grammatical categories can take a limited number of values—e.g., four for gender, three for number, a dozen for case—while tokens and lemmas may have thousands of values each. Statistical measures that work well for grammatical categories are not applicable to lexical units. For grammatical categories, we use the *frequency ratio*:

$$frequencyratio(c_i) = \frac{Q(c_i)}{P(c_i)}$$

If *frequency ratio* > 1, then the category's valueis selected by the pattern. For example, the frequency ratio for the pattern [*bez* 'without' + Noun] shows that among all values of *Case* only two are significant: the genitive (FR = 0.98/0.38 = 2.58) and the so-called "second" genitive (FR = 0.015/0.0019 = 7.89); thus, one of the main outputs of the system is that [*bez* 'without' + Noun.GEN] and [*bez* 'without' + Noun.GEN2] are stable *colligations*, which agrees with theory.

If lemmas or tokens receive a high Kullback-Leibler rank for the given pattern, the output should also include a list of stable *lexical items*; these are the main

focus of this article. The next step of the algorithm determines the particular lexical items that form stable *collocations* and the semantic classes (*constructions*). These two phenomena require different approaches, which are described in Sections 3 and 4, respectively.

3. Collocation extraction

3.1. Collocation extraction method

A variety of methods has been proposed to deal with collocation extraction. Pecina (2005) surveys 87 statistical measures and methods, and even that is not a complete list. The best known metrics are, for example, *Mutual Information* (MI; Church & Hanks 1990), the *Dice* coefficient (Daudaravicius 2010), *t-score* (Church et al. 1991) and *log-likelihood* (Dunning 1993). These are the most commonly used metrics for extraction of MWEs; however, they have some disadvantages. For example, MI tends to overestimate the low-frequency collocations, while the t-score mainly identifies high-frequency constructions. A choice of an appropriate statistical measure to rank MWEs depends on the goals and the data at hand. The current implementation of our algorithm calculates several measures for a given bigram and allows the human expert to rank bigrams based on various principles; these measures include both standard statistics, such as the t-score and MI, and several novel weighted schemes which we propose in our experiments.

In our previous study (Kormacheva et al. 2014), six statistical measures were compared and tested on 25 Russian prepositions; as a result, we came to the conclusion that the measure that best suits our goals is the *refined weighted frequency ratio (wFR)*. We use this measure in our work to determine fixed expressions.

Refined frequency ratio is calculated as follows:

$$FR(p,w) = \frac{f(p,w)}{f(w)}$$

where *p* is the pattern, *w* is the lexical unit that can match the pattern, *f (p, w)* is the absolute frequency of the lexical unit in the pattern, and *f (w)* is the absolute frequency of the lexical unit in the general corpus.

'Refined' here means that we distinguish between ambiguous grammatical forms present in the corpus. For example, it is typical in Russian that many tokens syncretize with others that belong to the same paradigm; for example, in the fixed expression *bez dela* 'at loose ends', *dela*, lit. 'business' is an ambiguous form, and the ordinary frequency ratio for this collocate would be *FR*=37/2146=0.017.

The refined frequency ratio allows us to give more weight to this collocation: *RFR*=37/1326=0.027.

Then the *weighted frequency ratio* is calculated as the refined frequency ratio multiplied by the logarithm of the unit frequency in the general corpus:
The idea behind this measure is as follows:

$$wFR(p,w) = FR(p,w) \times log\big(f(w)\big)$$

Let us consider two words, *w1*, which appears in the corpus 2 times, and *w2*, which appears in the corpus 1,000 times. Let *f(p, w1)* = 1, *f(p, w2)* = 500; hence, *FR(p, w1)* = *FR(p, w2)* = 0.5. It is obvious that *w1* may appear with the pattern accidentally, whereas the fact that *w2* occurs with the pattern 500 times out of 1,000 is meaningful. We multiply the frequency ratio by the logarithm of the word frequency to give more weight to frequent words.

Performance is evaluated using the uninterpolated average precision (UAP, Villada Moirón & Tiedemann 2006) which indirectly measures recall: at each point *c* of the ranking *r* where a relevant entryis found, the precisionis computed and all precision points are then averaged:

$$UAP(r) = \frac{\sum P(S_1 \ldots S_c)}{|S_c|}$$

3.2. Lemma collocations

This section focuses on the evaluation of the semantic stability of MWEs automatically obtained for the following patterns: [X.Adjective+Noun], [Adjective+X. Noun] and [Verb + X.Noun], where *X* is an unknown word belonging to a certain part of speech. We investigate only lemma collocations, putting token collocations aside, thus evaluating only those MWEs that are stable irrespective of the actual surface form; for example:

vysokij čelovek	*vysokogo čeloveka*	*vysokomu čeloveku,*	*etc.*
tall.NOM man.NOM	*tall.GEN man.GEN*	*tall.DAT man.DAT,*	*etc.*

For each pattern, ten of the most frequent lemmas were investigated in detail. The goal was to analyze the kinds of collocations that are extracted by our algorithm and whether it is possible to account for all extracted units. Owing to their statistical unreliability, words with the corpus frequency = 1 are filtered out. For each word, the obtained collocates were sorted according to the *wFR* (see Section 3.1.),

and the top 100 collocations were manually annotated as either free or fixed (by fixed, we mean here collocations in which the meaning is either bound by the context or is non-compositional).

However, since many extracted MWEs do not fulfill the requirement of idiomaticity (non-compositional meaning), we have also concentrated on the analysis of non-idiomatic collocations. We turned our attention to *constructions* in which a certain collocation is one among many representatives. We assume that constructions have potential as a theoretical background in building semantic groups of collocations by capturing a lexeme's semantic preferences. Where possible, the obtained collocates were divided into classes that represent lexemes with similar semantic properties, designating, for example, 'nationality', 'food', 'furniture', and so on. Such collocations are not necessarily semantically non-compositional, but include words that share grammatical and semantic features and thus belong to the constructions (a syntactic phrase in which at least one position is occupied by a word-form belonging to a certain semantic class). We illustrate this with six examples, where bases of the collocations are a frequently used adjective or a verb, and each syntactic group consists of the given word and a list of variables with similar meaning. The reason behind our choice is that collocations crucially depend on the size and representativeness of a corpus. As a result, many of the presented collocations hinge on frequent words that have very general semantics, such as parametric adjectives or lexical functional verbs. The chosen case studies illustrate different ratios of fixed expressions and constructions in these MWEs.

(1) Case study "*vysokij* 'high/tall' + Noun"[4]

Among the top 100 collocates extracted for the query "*vysokij* 'high/tall' + Noun", 26 were manually attested[5] as being included in fixed expressions (*vysokaja tehnologija* 'high technology/high tech', *vysokaja* častota 'high frequency', *vysokoe davlenie* 'high/heavy pressure', etc.) and 16 were considered to be part of three different constructions (the last implies a minimum of three lemmas from the top 100). The UAP of 0.51 (see Table 1) means that relevant collocates are distributed equally in the top 100 (ranking by wFR), although they are slightly more prevalent among those having the highest wFR. The three constructions are:

4 Russian is a free word order language. The word order hereinafter corresponds to that in the collocations analyzed. E.g., *vysokaja rentabel'nost'* 'high profitability' is taken into account, but not *rentabel'nost' vysoka* 'profitability [is] high', just as *polučil Vladimir* lit. 'has got Vladimir', but not *Vladimitr pulochil 'Vladimir has got'*.

5 We have also looked them up in dictionaries of Russian idioms.

6 MWEs: [*vysokij* 'high/tall' + characteristics of materials/products]: *pročnost'* 'durability', *iznosostojkost'* 'wear resistance', *pomehozaščiščennost'* 'noise-immunity', etc.;

5 MWEs: [*vysokij* 'high/tall' + vocal range]: *tenor* 'tenor', *mecco-soprano* 'mezzo-soprano', *tessitura* 'tessitura', etc.

5 MWEs: [*vysokij* 'high/tall' + economic terms]: *likvidnost'* 'liquidity', *rentabel'nost'* 'profitability', *dividend* 'dividend', etc.

(2) Case study "*molodoj* 'young' + Noun"

In this query, an even greater number of collocates can be accounted for by constructional preferences. Fifty-six nouns returned for this query can be described as falling into one of the following six constructions:

22 MWEs: [*molodoj* 'young' + name/surname]: *Gertruda* 'Gertruda', *Irma* 'Irma', *Doronin* 'Doronin', etc.;

19 MWEs: [*molodoj* 'young' + profession]: *pevec* 'singer', *žurnalistka* 'female journalist', *atlet* 'athlete', etc.;

5 MWEs: [*molodoj* 'young' + nationality]: *kazah* 'Kazakh', *vengerec* 'Hungarian', *afrikanec* 'African', *bel'giec* 'Belgian', *britanec* 'Briton';

4 MWEs: [*molodoj* 'young' + animal (in diminutive form)]: *burunduček* 'chipmunk', *l'venok* 'lion', *lisička* 'fox', *barašek* 'lamb';

3 MWEs: [*molodoj* 'young' + bird]: *čibis* 'lapwing', *bekas* 'snipe', *čiž* 'siskin';

3 MWEs: [*molodoj* 'young' + type of forest]: *osinnik* 'aspen forest', *sosnjak* 'pine forest', *el'nik* 'fir forest'.

These constructions constitute a considerable portion of the extracted MWEs; six constructions account for 56 of 100 MWEs (see Table 1, column 4). Adding three fixed expressions (*molodaja gvardija* 'young guard', *molodoe pokolenie* 'young generation', *molodoj čelovek* 'young man'), we get 59 MWEs that can be explained either as a collocation or as a representation of the constructions.

(3) Case study "Adj + *mesto* 'place'"

This query uses as an example, in which the number of fixed expressions is far greater than the number of words belonging to the constructions.

Fixed expressions (26): *othožij* (‘latrine’), *prizovoj* lit. ‘prize (adj.)’ (‘prize/prize-winning place’), *ukromnyj* lit. ‘secluded’ (‘ivy-bush’), *zlačnyj* (‘tenderloin/hot spot’), *početnij* lit. ‘honored’ (‘place of pride’), *rabočij* (‘workplace’), *vidnyj* lit. ‘visible/conspicuous’, *posadočnyj* (‘seat’), *bol’noj* lit. ‘sore’ (‘sore point’), *spal’nyj* lit. ‘sleeping’ (‘berth’), etc.

Constructions (9 MWEs):
3 MWEs: [positive personal attitude + *mesto* ‘place’]: *ljubimyj* ‘favorite’, *lučšyj* ‘the best’, *izljublennyj* ‘favorite’;

3 MWEs: [type of relief + *mesto* ‘place’]: *vozvyšennyj* ‘high/elevated’, *goristyj* ‘mountainous’, *nizmennyj* ‘low-lying’;

3 MWEs: [location near the sea + *mesto* ‘place’]: *pribrežnij* ‘coastal’, *černomorskij* ‘Black Sea (adj.)’, *kurortnyj* ‘resort’.

(4) Case study “Adj + *vopros* ‘question/issue’”

Fixed expressions (16): *nemoj* lit. ‘dumb’ (‘unspoken’), *kaverznyj* ‘tricky’, *večnyj* ‘eternal’, žiznennyj lit. ‘vital’ (‘problem of life’), *spornyj* lit. ‘disputable’ (‘moot point/point at issue’), *molčalivyj* lit. ‘taciturn/silent’ (‘unspoken’), *ščekotlivyj* lit. ‘ticklish’ (‘delicate question/ticklish problem’), etc.

Constructions (25 MWEs):
12 MWEs: [of great importance + *vopros* ‘question/issue’]: *žiznennyj* ‘vital’, *principial’nyj* ‘fundamental’, *važnejšij* ‘the most important’, *kardinal’nyj* ‘cardinal/fundamental’, *fundamental’nyj* ‘fundamental’, *aktual’nyj* ‘actual’, *sud’bonosnyj* ‘determining/crucial’, *smysložiznennyj* ‘vital’, *neotložnyj* ‘pressing/urgent’, *večnyj* ‘eternal’, *glavnyj* ‘the most important’, *volnujuščij* ‘exciting’, *bazovyj* ‘base’;

9 MWEs: [quality of + *vopros* ‘question’]: *nevnjatnyj* ‘indistinct’, *kosvennyj* ‘indirect’, *otvlečennyj* ‘abstract’, *abstraktnyj* ‘abstract’, *idiotskij* ‘stupid’, *primitivnyj* ‘primitive’, *prosten’kij* ‘simple’, *nelepyj* ‘absurd’, *durackij* ‘stupid’;

3 MWEs: [habitat + *vopros* ‘issue’]: *territorial’nyj* ‘territorial’, *kvartirnyj* ‘housing’, *žiliščnyj* ‘housing’ (the last two meaning ‘housing issue’);

This query gives a clear example of a case in which most stable, idiomatic expressions not only constitute the ultimate level of stableness, but also are simultaneously representatives of certain constructions, generalized on a higher level

of stability. For example, adjectives in the idiomatic expressions *večnyj vopros* 'eternal question' and žiznennyj *vopros* 'a life-and-death question' also belong to the group of collocates like *važnejšij* 'the most important', *principial'nij* 'principal', *aktual'nyj* 'actual' and others with the common meaning of 'being important'.

(5) Case study "*polučit'* 'to get/to receive' + Noun"

Although no expressions were tagged as fixed for this query, the question actively participates in various constructions. Among the top 100 collocates, the following ten constructions can be distinguished:

6 MWEs: [*polučit'* 'to get/to receive' + message]: *izveščenie* 'notification', *poslanie* bazovyj 'message', *posylka* 'package/parcel', *telegramma* 'telegram', *pis'mo* 'letter', *zapiska* 'note/slip';

6 MWEs: [*polučit'* 'to get/to receive' + proper name]: *Gil'ermo*, *Šerlok*, *Mefistofel'*, *Hripušin, Vladimir, Pavel*;

4 MWEs: [*polučit'* 'to get/to receive' + document]: *attestat* 'certificate (usually for education)', *diplom* 'diploma', *udostoverenie* 'certificate', *licenzija* 'license';

3 MWEs: [*polučit'* 'to get/to receive' + injury]: *ranenie* 'wound', *travma* 'trauma', *sotrjasenie* 'concussion';

3 MWEs: [*polučit'* 'to get/to receive' + a large amount]: *tonna* 'ton', *massa* 'mass', *kuča* 'piles (of)';

3 MWEs: [*polučit'* 'to get/to receive' + name']: *imja* 'name (usually human)', *nazvanie* 'name', *prozviščе* 'nickname';

3 MWEs: [*polučit'* 'to get/to receive' + order]: *prikazanie* 'order/command', *rasporjaženie* 'order/instruction', *prikaz* 'order/command';

3 MWEs: [*polučit'* 'to get/to receive' + vaccination]: *vakcina* 'vaccine', *privivka* 'inoculation', *in"ekcija* 'injection';

3 MWEs: [*polučit'* 'to get/to receive' + title]: *titul* 'title', *stepen'* 'degree', *zvanie* 'title/status';

3 MWEs: [*polučit'* 'to get/to receive' + positive feedback]: *podderžka* 'support', *blagodarnost'* 'gratitude', *odobrenie* 'approval'.

(6) Case study "*javljat'sja* 'to be/to serve (as)' + Noun"

Finally, there are words that tend to be neither parts of fixed expressions nor do they participate in constructions. The manual analysis of the obtained collocations showed no relevant results for this query. The obtained collocates include such words as *sponsor* 'sponsor', *avtomatizacija* 'automation', *poza* 'posture', *istočnik* 'source', and *vozbuditel'* 'stimulus' among others.

Similar observations can be made for other queries. In all cases, taking constructions into account significantly improves the results. We propose that such constructional preferences can be found automatically, and the next section is dedicated to this topic. In Table 1, we present the manual evaluation of the number of fixed expressions and the number of words participating in various constructions for frequently used Russian words. To construct queries, we have selected the most frequent nouns, adjectives, and verbs and extracted the most stable lemmas that co-occur with them. For each query, the first 100 bigrams were sorted according to the weighted frequency ratio (if the output contained less than 100 words, all results are taken) and manually grouped into semantic classes that represent certain constructions. The uninterpolated average precision (UAP) is used to evaluate the results; it reflects the total number of relevant results for a given query (where 1 means that all results in a query fall into either fixed expressions or constructions and 0 means that there are no fixed expressions in the response). The data, although not without gaps, show the benefit of constructions in describing automatically obtained MWEs.

To sum up, using this method we obtain MWEs of different kinds. They include lexically stable expressions, which are collocations in a narrow sense, as well as expressions constrained on the semantic class level, which are constructions. On average, among the top 100 MWEs in the examined queries only 7.5 percent are purely idiomatic collocations. Other MWEs are frequently used and stable, but not idiomatic. So the question then arises, with this algorithm is it possible to extract mere collocations? The answer to this question is no, and below we will analyze the main factors that result in this answer.

Table 1. The number of fixed expressions (expres.) and constructions (constr.) for frequent Russian bigrams.

Query	expres.	constr.	UAP	Query	expres.	constr.	UAP
molodoj 'young' + N	3	56	0.69	A + *vremja* 'time'	13	46	0.67
horošij 'good' + N	6	34	0.36	A + *god* 'year'	10	41	0.66
ravnyj 'equal' + N	3	27	0.35	A + *čelovek* 'human'	4	33	0.49
vysokij 'high' + N	26	16	0.51	A + *vopros* 'question'	16	25	0.44
poslednij 'last' + N	12	5	0.23	A + *delo* 'matter'	9	24	0.48
krajnij 'extreme' + N	7	31	0.43	A + *den'* 'day'	11	40	0.69
glavnyj 'main' + N	9	25	0.46	A + *žizn'* 'life'	3	6	0.2
malen'kij 'small' + N	0	28	0.37	A + *mesto* 'place'	26	9	0.54
raznyj 'different' + N	2	5	0.07	A + *rabota* 'work'	20	29	0.57
važnyj 'important' + N	8	19	0.31	A + *slučaj* 'case'	10	35	0.6
videt' 'to see' + N	0	42	0.41				
delat' 'to do' + N	14	12	0.31				
znat' 'to know' + N	0	32	0.29				
imet' 'to have' + N	14	6	0.16				
sdelat' 'to do' + N	7	17	0.26				
stat' 'to become' + N	0	39	0.3				
javljat'sja 'to be' + N	0	0	0				
idti 'to go' + N	0	30	0.35				
dat' 'to give' + N	24	27	0.57				
polučit' 'to get/to receive'+ N	0	37	0.53				
				Average			**0.41**

The first reason is that the lack of stable expressions in a query output is due to the low value of the weighted frequency ratio for a lemma/token. It means that a given word does not tend to coin fixed expressions, and what we get is a rather unstable list of lemmas/tokens that are more or less frequently used, but do not have any idiosyncratic features.

The occurrence of irrelevant items among the results is also due to a substantial number of words that have low corpus frequency, which significantly degrades the statistics, as is usually the case with small numbers. They, however, cannot be filtered out, because the small corpus size inevitably implies that some relevant cases will occur rarely as well. This is one point to be improved in the future by expanding the corpus size.

To conclude, the top 100 stable MWEs is not just a list of idioms and the like. An idiomatic nature is but one of the many reasons why certain lemmas or tokens

tend to stay together. Another reason is a semantic cohesion of word classes with similar meaning. We propose that such constructions can be extracted automatically, which we will demonstrate in the next section.

4. Toward automatic construction extraction

In theory, we assume that semantic preferences of the input patterns can be found using algorithms similar to what has been established in the previous sections. If the corpus were semantically annotated, then we would be able to group words according to their semantic tags (e.g., 'animal', 'nationality', etc.) and to extract different kinds of constructions in the same way as we do with grammatical categories or lemmas. Unfortunately, we do not have access to Russian data suitable for this task, nor are semantically annotated Russian thesauri (e.g., a Russian WordNet) available. However, we can try to *bootstrap* semantic classes from the data using *distributional analysis*.

The core idea of this approach is that semantic word similarity is related to the distributional properties of the context. The underlying statement—"You shall know a word by the company it keeps!" (Firth 1957: 11)—has a long history in linguistics, but it became especially popular in recent decades when advanced methods in machine learning and statistics have allowed researchers to study distributional preferences experimentally on a *corpus scale* (Baroni et al. 2014; Huang et al. 2012; Van de Cruys 2010; Erk & Padó 2008).

The most recent burst in this line of research has been thanks to the development of a novel machine learning technique called *neural networks*. Using this approach, Tomas Mikolov and his colleagues have developed a *word2vec* tool (Mikolov et al. 2013a, 2013b) suitable for building neural-network models for text data. The tool constructs a vector representation of a given corpus. Words that have similar distribution patterns in the corpus are close to each other in the vector space. The crucial point here is that words with similar distribution patterns have similar meanings; thus, the distributional similarity is interpretable as a semantic distance. The tool also provides an effective way to investigate further the calculated semantic space; it can—*inter alia*—be used for searching semantic distances between given words or to list words that are most similar to a given one.

The general interest in distributional semantics has resulted in a number of studies that apply this approach to the Russian language. A shared task on semantic word similarity for Russian was organized in 2015 (Panchenko et al. 2015), where one of the best scores demonstrated the RusVectores project (Kutuzov & Andreev 2015), which applied *word2vec* to the Russian data. The RusVectores application is freely available at ling.go.mail.ru/dsm/en, which allowed us to use the model as

an initial step in automatic construction extracting. One more advantage, and by no means the least, is that this application is based on the same data that we used in our research (Russian National Corpus).

To carry out comparable research, we based the initial word lists for each query on the same lists that we used for manual evaluation (the top 100 lemmas ranked by *wFR*; see Section 3). They were grouped into semantic clusters using the following formal procedure:

1. The first word in the list (the word with the highest wFR) served as a seed for the first semantic cluster;
2. Taking this word, we calculated the distance between the seed and all other words in the list using the *word2vec* model. If the distance appeared to be under a certain threshold,[6] then the words were grouped together. As soon as the first cluster was formed, all words in the cluster were excluded from further processing.
3. We then selected the second seed, which is the word with the highest wFR among those not yet clustered. The second cluster was formed from this seed word and all remaining words if their distance from the seed was below the same threshold.
4. This procedure was repeated until all words were grouped into clusters. The clusters containing one or two words were excluded from analysis; the clusters that consisted of three or more words were considered to represent certain constructions.

In comparing calculated groups with manually attested constructions in Section 3, we learned — to our pleasure— that the algorithm extracts quite comparable lists of words. Below, we consider several examples (both good and somewhat strange) of automatically found constructions (words that correspond to those manually analyzed in Section 3 are in bold-face type).

(1) "*polučit'* 'to get/to receive' + Noun"

a. *sotrjasenie* 'concussion', *travma* 'trauma', *ranenie* 'wound'

b. *attestat* 'certificate (usually for education)', *diplom* 'diploma', *udostoverenie* 'certificate'

6 The particular value of the threshold is a subject of our future research; in this paper, words with a semantic distance of less than 0.3 were grouped.

c. *prikazanie* 'order/command', *rasporjaženie* 'order/instruction', *prikaz* 'order/command';

d. *prozvišče* 'nickname', *nazvanie* 'name', *imja* 'name (usually human)'

e. *poslanie* 'message', *telegramma* 'telegram', *pis'mo* 'letter'

(2) "*vysokij* 'high/tall' + Noun"

a. *kabluk* 'heel', *šnurovka* 'lacing', *botfort* 'jackboot', *ščikolotka* 'ankle'

b. *pomehozaščiščennost'* 'noise-immunity', *iznosostojkost'* 'enduring quality', *proizvoditel'nost'* 'productivity', *bystrodejstvie* 'promptitude', *čuvstvitel'nost'* 'sensitiveness'

c. *uroven'* 'level', *rang* 'rank', *stepen'* 'degree', *koncentracija* 'concentration', *pokazatel'* 'index'

d. *ekonomičnost'* 'economy', *prohodimost'* 'passability', *pročnost'* 'durability', *èffektivnost'* 'effectiveness', *mobil'nost'* 'mobility', *točnost'* 'precision', *četkost'* 'accuracy/clearness'

e. *marža* 'margin', *likvidnost'* 'liquidity', *rentabel'nost'* 'profitability', *dohodnost'* 'profitableness'

f. *temperatura* 'temperature', *davlenie* 'pressure', *naprjaženie* 'voltage'

However, in some cases automatic clustering is too split up. For example, the case [*molodoj* 'young' + Noun] below gives 19 names of professions separated into three groups (a, b, and c):

(3) "*molodoj* 'young' + Noun"

a. *pevec* 'singer', *vokalist* 'vocalist', *solistka* 'soloist (fem.)', *pevica* 'singer (fem.)'

b. *figurist* 'figure-skater', *biatlonistka* 'biathlete (fem.)', *tennisist* 'tennis-player', *atlet* 'athlete'

c. *reformator* 'reformer', *literator* 'writer', *učenyj* 'scientist', *fizik* 'physicist'

d. *sosnjak* 'pinery', *el'nik* 'spruce forest', *perelesok* 'coppice',

e. *čelovek* 'man', *paren'* 'guy', *ženščina* 'woman'

Intuitively, the groups *a, b*, and c should be grouped together, as we have done in

Section 3, but the algorithm splits them up. This result cannot be considered completely wrong, since these groups have more specific meanings, such as 'singers' or 'sportsmen/sportswomen'. It depends on the threshold used to group the words; the higher this threshold, the more semantically specific is the cluster, but the lower the threshold, the noisier are the results.

We also learned that there is variation between different queries. For example, in case (4), Adj + *vorpos* 'question/issue', the clusters obtained are more questionable because the noun is semantically ambiguous. The algorithm groups collocates relating to 'question' {*kaverznyj* 'tricky'*, ritoričeskij* 'rhetorical'*, nedoumennyj* 'puzzled'*, rezonnyj* 'reasonable'*, ehidnyj* 'retortive'*,* etc.} together with those connected to 'issue' {*nerazrešimyj* 'insolvable'*, sud'bonosnyj* 'determining/crucial'*, nerešennyj* 'unsolved'*, delikatnyj* 'delicate'*,* etc.}.

(4): Case study "Adj + *vorpos* 'question/issue'

a. *kaverznyj* 'tricky', *ritoričeskij* 'rhetorical', *spornyj* 'disputable', *nedoumennyj* 'puzzled', *nerazrešimyj* 'insolvable', *sud'bonosnyj* 'determining/crucial', *rezonnyj* 'reasonable', *ehidnyj* 'venomous', *nerešennyj* 'unsolved', *delikatnyj* 'delicate', *ščekotlivyj* 'ticklish', *nelepyj* 'absurd', *durackij* 'stupid', *trudnyj* 'difficult', *složnyj* 'hard', *idiotskij* 'idiotic'

b. *agrarnyj* 'agrarian', *žiliščnyj* 'housing', *social'no-èkonomičeskij* 'social-economic', *social'nyj* 'social', *političeskij* 'politic'

c. *neskromnyj* 'indelicate', *nazojlivyj* 'importunate', *nepriličnyj* 'indecent', *prazdnyj* 'idle'

d. *zakonomernyj* 'appropriate', *estestvennyj* 'natural', *neizbežnyj* 'unavoidable'

e. *èkstravagantnyj* 'extravagant', *pošlyj* 'vulgar', *prosten'kij* 'simple', *primitivnyj* 'primitive'

f. *aktual'nyj* 'topical', *fundamental'nyj* 'fundamental', *volnujuščij* 'exciting'

g. *taktičeskij* 'tactical', *organizacionnyj* 'organizational', *metodologičeskij* 'methodical', *praktičeskij* 'practical', *teoretičeskij* 'theoretical', *tehničeskij* 'technical', *operativnyj* 'operational'

h. *filologičeskij* 'philological', *filosofskij* 'philosophical', *bogoslovskij* 'theological', *lingvističeskij* 'linguistic'

i. *nevnjatnyj* 'indistinct', *toroplivyj* 'hasty', *nastojčivyj* 'insistent', *neuverennyj* 'uncertain' etc.

5. Conclusion

In grammar books, language is often described as an efficient system of rules operating continuously on each level, and the 'exceptions' that destabilize the language system. However, we argue that the MWEs seem to be an extremely important part of language usage and are hard to pinpoint on a certain language level. Some MWEs are less frequent and may be dropped away, while others will—in all probability—be crystallized into new rule-driven structures. MWEs are inter level and exceptional by nature; however, this does not make it impossible to formulate generalizations about their idiosyncrasies. These generalizations are probabilistic rather than rule-based.

> By adopting a theory in which lexical items above the word level can license syntactic structures, we can incorporate clichés and idioms into the lexicon. Furthermore, many of the curious properties of idioms are altogether parallel to those already found within lexical word grammar, so they add no further complexity to grammatical theory. (Jackendoff 1995:153)

The basic idea behind our algorithm is to locate MWEs of different kinds in a unified fashion. The algorithm predicts the most stable features in an n-gram, a linear sequence of tokens, where these features may be morphological, lexical, syntactic, or semantic.

In this article, we took a closer look at lexical combinations in the Russian language that are traditionally treated as collocations. We, however, showed that there are two different types of lexically restricted MWEs. The first is indeed collocations, where a lemma/token chain predicts the next/previous lemma or token. The ultimate example of this kind of lexically stable chain is an idiom about which no generalization is possible, e.g., *lo and behold*. However, some collocations appeared to be stable on a more abstract level of generalization. They represent constructions in which lexical variables are replaceable, but belong to the same semantic class, e.g., *sleight of [hand/mouth/mind]*. In this case, even if a specific collocation as such is rare, prediction of the whole semantic class is highly possible. It is worth pointing out that, formally, there is no border between collocations and constructions, and any collocation can be viewed as a construction with one or more lexical values. The analysis presented in Section 3 confirms this idea. As for automatically defining semantic classes, we see two ways to do this. The first is by using a semantic annotation given in a corpus that seems to be promising, yet is unrealistic owing to the lack of publicly available annotation. The alternative is to use methods of distributive semantics in order to find semantic classes automati-

cally. We have made the first step in this direction in the present article, and we will concentrate on taking this further in the future.

Acknowledgments

Part of this work was supported by grant from the Research Foundation of the University of Helsinki. We would also like to thank R. Yangarber and M. Pierce for their academic and technical support, as well as the RNC developers (especially E. Rakhilina and O. Lyashevskaya) and S. Sharoff for sharing their data.

References

Baroni, Marco, Raffaella Bernardi & Roberto Zamparelli 2014. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology* 9: 241–346.

Church, Kenneth W., William Gale, Patrick Hanks & Donald Kindle 1991. Using statistics in lexical analysis. *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*, ed. Uri Zernik, 115–164. Hillsdale, New Jersey/Hove, UK: Lawrence Erlbaum.

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

Daudaravicius, Vidas 2010. Automatic identification of lexical units. *Informatica: An International Journal of Computing and Informatics* 34(1): 85–91.

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

Erk, Katrin & Sebastian Padó 2008. A structured vector space model for word meaning in context. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, eds. M. Lapata & H. T. Ng, 897–906. Honolulu, HI: Association for Computational Linguistics.

Fillmore, Charles, Paul Kay & Catherine O'Connor 1988. Regularity and idiomaticity in grammatical constructions: the case of *Let alone*. *Language* 64(3): 501–538.

Firth, John R. 1957. A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis*, ed. John R. Firth, 1–32. Oxford: Philological Society. [Reprinted in *Selected Papers of J.R. Firth 1952-1959*, ed. Frank R. Palmer, 1968. London: Longman.]

Goldberg, Adele E. 1996. Construction grammar. *Concise Encyclopedia of Syntactic Theories*, eds. E. Keith Brown & Jim E. Miller, 68–70. Oxford: Elsevier Science Ltd.

Gries, Stefan Th. & Anatol Stefanowitsch 2004. Extending collostructional analysis: A corpus-based perspective on alternations. *International Journal of Corpus Linguistics* 9(1): 97–129.

Hoey, Michael 2004. The textual priming of lexis. *Corpora and Language Learners*, eds.

Guy Aston, Silvia Bernardini & Dominic Stewart, 21–41. Amsterdam: John Benjamins.

Huang, Eric H., Richard Socher, Christopher D. Manning & Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics:* Long Papers-Volume 1, 873-882. Stroudsbrug, PA: Association for Computational Linguistics.

Hunston, Susan & Gill Francis 2000. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.

Jackendoff, Ray 1995. The Boundaries of the Lexicon. *Idioms: Structural and Psychological Perspectives*, eds. Martin Everaert, Erik-Jan van der Linden, André Schenk & Rob Schreuder, 133–165. Hillsdale, New Jersey/Hove, UK: Lawrence Erlbaum.

Kopotev, Mikhail, Lidia Pivovarova, Natalia Kochetkova & Roman Yangarber 2013. Automatic detection of stable grammatical features in n-grams. *Papers from the 9th Workshop on Multiword Expressions (MWE 2013). Workshop at NAACL 2013,* 73–81. Atlanta, Georgia: The Association for Computational Linguistics.

Kormacheva, Daria, Lidia Pivovarova & Mikhail Kopotev 2014. Automatic collocation extraction and classification of automatically obtained bigrams. *Workshop on Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations (CCLCC 2014, Tübingen, Germany)*, eds., Verena Henrich & Erhard Hinrichs, 27–33. Tübingen: University of Tübingen.

Kutuzov, Andrey & Igor Andreev 2015. Texts in, meaning out: neural language models in semantic similarity task for Russian. *Proceedings of the Dialog 2015 Conference, Moscow, Russia*. Available at: http://arxiv.org/abs/1504.08183.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean 2013a. Efficient estimation of word representations in vector space. *Proceedings of Workshop at International Conference on Learning Representations (ICLR)* 2013. Available at: http://arxiv.org/abs/1301.3781.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado & Jeffrey Dean 2013b. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* 26 (*Proceedings of NIPS*, 2013), eds. C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K.Q. Weinberger, 1–9.

Moon, Rosamund 1998. *Fixed Expressions and Idioms in English: A Corpus-based Approach*. Oxford: Clarendon Press.

Panchenko, A., N. V. Loukachevitch, D. Ustalov, D. Paperno, C. M. Meyer, & N. Konstantinova 2015. Russe: The first workshop on Russian semantic similarity. *Proceeding of the Dialogue 2015 Conference*, 89–106.

Pecina, Pavel 2005. An extensive empirical study of collocation extraction methods. *Proceedings of the ACL Student Research Workshop*, 13–18.

Rakhilina, Ekaterina V. (ed.) 2005. *Natsionalny korpus russkogo jazyka: 2003—2005*. Moskow: Indrik.

Sharoff, Serge & Joakim Nivre 2011. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. *Kompojuternaja lingvistika i intellektualnye tekhnologii*, 591-604. Moscow. Available at: http://corpus.leeds.ac.uk/serge/publications/2011-dialog.pdf.

Sinclair, John McH. & Anna Mauranen 2006. *Linear Unit Grammar: Integrating Speech and Writing*. Amsterdam: John Benjamins.

Van de Cruys, Tim 2010. A non-negative tensor factorization model for selectional preference induction. *Natural Language Engineering* 16(4): 417–437.

Villada Moirón, Begoña & Jörg Tiedemann 2006. Identifying idiomatic expressions using automatic word-alignment. *Proceedings of the EACL 2006 Workshop on Multi-word Expressions in a Multilingual Context*, 33–40.