CoCoCo. automatic extraction of Russian collocations, colligations, and constructions

Lidia Pivovarova, Mikhail Kopotev, Daria Kormacheva,

University of Helsinki
CoCoCo

Collocations, Colligations & Corpora project aims to develop methods for extraction, classification and analysis of multi-word expressions (MWEs).

- University of Helsinki, team-leader M. Kopotev
Motivation: grammatical profiling

(Gries, Divjak (2009); Gries (2010); Janda, Lyashevskaya (2011); Divjak, Arppe (2013))

Grammatical profile – distribution of grammatical and lexical features of the context, which are relevant for a particular word class.

Main difference: profiles are extracted from corpus rather than set a priori

Automatic determination of words’ distributional preferences:

- Implementation of the model able to process MWEs of various nature on an equal basis
- The model compares the strength of various relations between the tokens in a given n-gram and searches for the “underlying cause” that binds the words together, whether it is lexical, grammatical, or a combination of both
- Developing an application for people studying foreign languages
What do we get from extracting MWEs?

• grammatically restricted colligations: try to + V.Inf
• collocations (incl. idioms): lo and behold
• semantic constructions: sleight of [hand/mouth/mind]
What do we get from extracting MWEs?

- GRET’ ‘warm (up)/ heat (up)’ + N
- DUŠU ‘soul’
- KROV’ ‘blood’
- VODU ‘water’
- MOLOKO ‘milk’
- ČAJ ‘tea’
- RUKI ‘hands’
- LADONI ‘palms’
- NOGI ‘feet’
- KOPYTA ‘hoofs’
- SPINU ‘back’
- MAŠINU ‘car’
- MOTOR ‘motor’
What do we get from extracting MWEs?

Generalization about automatically extracted Russian collocations

What do we get from extracting MWEs?

Colligations

Colligation – the grammatical company a word keeps (or avoids keeping) and the positions it prefers.

(Hoey, 2004)
What do we get from extracting MWEs?

Collocations typically denote frequently repeated or statistically significant co-occurrences, whether or not there are special semantic bonds between collocating items.

(Moon, 1998)
What do we get from extracting MWEs?

GRET’ ‘warm (up)/ heat (up)’ + N

DUŠU ‘soul’
KROV’ ‘blood’
VODU ‘water’
MOLOKO ‘milk’
ČAJ ‘tea’
RUKI ‘hands’
LADONI ‘palms’
NOGI ‘feet’
KOPYTA ‘hoofs’
SPINU ‘back’
MAŠINU ‘car’
MOTOR ‘motor’

Constructions
Construction – 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)
Algorithm

For each part of speech:

- Stable features
  - For each grammatical feature:
    - particular values for the features
    - most specific tokens / lemmas
    - most specific semantic classes

Output:

- Colligations
- Collocations
- Construction
Algorithm

For each part of speech:

Stable features

For each grammatical feature:

- particular values for the features
- most specific tokens / lemmas
- most specific semantic classes

Output:

- Colligations
- Collocations
- Construction s
Kullback-Leibler divergence

\[ D_{KL}(Q_C || P_C) = \sum_{i=1}^{N} Q(c_i) \times \log \frac{Q(c_i)}{P(c_i)} \]

Kopotev et al. 2013

Generalization about automatically extracted Russian collocations
Generalization about automatically extracted Russian collocations

Algorithm

For each part of speech:

- For each grammatical feature:
  - particular values for the features
  - most specific tokens / lemmas
  - most specific semantic classes

Output:

- Colligations
- Collocations
- Constructions
Weighted frequency ratio

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

- Kopotev et al. 2013: research on bigrams beginning with prepositions; disambiguated subcorpus of RNC (a. 6 millions)
- Case category has the maximum \( D_{KL} \) for all the prepositions
- FR predicts the correct case with a precision of 95% and recall of 89%

\[ wFR(p, w) = FR(p, w) \times \log f(w) \]

- Kormacheva et al. 2014: research on bigrams matching the [Preposition + x.Noun] pattern; disambiguated subcorpus of RNC (a. 6 millions)
- Comparison of 6 evaluation measures (FR, wFR, MI, dice, t-score, frequency) for collocation extraction; \( wFR \) shows the best results
- The accuracy for different prepositions varies significantly – between 4% and 73%
Error analysis

Collocations:
- *bez pamjati* (without.PREP memory.NOUN.SG.GEN, 'like mad', 'passionately')
- *bez ceremonij* (without.PREP ceremony.NOUN.PL.GEN, 'informally')
- *u istokov* (at.PREP river source.NOUN.PL.GEN, 'at the origins')

<table>
<thead>
<tr>
<th>Preposition</th>
<th>f</th>
<th>rFR</th>
<th>wFR</th>
<th>MI</th>
<th>Dice</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bez (‘Without’)</td>
<td>72.86</td>
<td>68.38</td>
<td><strong>73.34</strong></td>
<td>7.17</td>
<td>5.83</td>
<td>72.60</td>
</tr>
<tr>
<td>U (‘Near/ At’)</td>
<td>3.97</td>
<td>1.92</td>
<td><strong>4.17</strong></td>
<td>0.00</td>
<td>0.00</td>
<td>2.92</td>
</tr>
</tbody>
</table>
Error analysis of $u$ ('near/ at')

- Constructions constitute a considerable part of the extracted bigrams:
  - 16: [$u$ 'near/at' + PART OF HOUSE]: okno 'window', kryl’ko 'porch', stena 'wall', etc.;
  - 13: [$u$ 'near/at/-' + ANIMAL]: koška 'cat', korova 'cow', mlekovitajuščee 'mammal', etc.;
  - 10: [$u$ 'near/at/-' + RELATIVE]: rebenok 'child', papa 'dad', tešča 'mother in law', etc.;
  - 8: [$u$ 'near/at' + PART OF INTERIOR]: stojka 'counter', televizor 'tv-set', kamin 'fireplace', etc.;
  - 6: [$u$ 'near/at/-' + NATIONALITY]: nemec 'German', russkij ‘Russian', cygan ‘Gypsy', etc.;

- Counting these bigrams as relevant collocations would increase the result from 4.17 to 73.82%
Algorithm

For each part of speech:

- For each grammatical feature:
  - particular values for the features
  - most specific tokens / lemmas
  - most specific semantic classes

Output:

- Colligations
- Collocations
- Constructions
Semantic clustering method

“You shall know a word by the company it keeps” (Firth, 1957)

Distributional semantics:
the semantic word similarity correlates with the
distributional properties of the context

• collecting contexts for each word in the corpus;
• obtaining pairwise semantic similarity between words;
• grouping words in semantic clusters.
Constructional profile for [molodoj ‘young’ + X]
Constructional profile for [molodoj ‘young’ + X]

- spruce
- birch
- poplar
- tree
- grass

- doctor
- professor
- director
Generalization about automatically extracted Russian collocations

BEZ ‘without’ + N

- PAMJATI ‘memory’
- CEREMONIJ ‘ceremony’
- GALSTUKA ‘tie’
- PERČATOK ‘gloves’
- POGON ‘epaulette’
- ŠAPKI ‘cap’
Generalization about automatically extracted Russian collocations

BEZ ‘without’ + N

PAMJATI ‘memory’

CEREMONIJ ‘ceremony’

GALSTUKA ‘tie’

PERČATOK ‘gloves’

POGON ‘epaulette’

ŠAPKI ‘cap’
Generalization about automatically extracted Russian collocations

BEZ ‘without’ + N

PAMJATI ‘memory’

CEREMONIJ ‘ceremony’

GALSTUKA ‘tie’

PERČATOK ‘gloves’

POGON ‘epaulette’

ŠAPKI ‘cap’
Generalization about automatically extracted Russian collocations

- BEZ ‘without’ + N
- PAMJATI ‘memory’
- CEREMONIJ ‘ceremony’
- GALSTUKA ‘tie’
- PERČATOK ‘gloves’
- POGON ‘epaulette’
- ŠAPKI ‘cap’

‘like mad, passionately’
‘informally’
‘informally’
The method extracts MWEs of different nature: collocations, colligations, constructions.

Most of the extracted MWEs are stable and frequently used, however not idiomatic.

Some part of the extracted bigrams can be described in terms of constructions that predict some grammatical and semantic features of a word class.
• The method extracts MWEs of different nature: collocations, colligations, constructions
• Most of the extracted MWEs are stable and frequently used, however not idiomatic
• Some part of the extracted bigrams can be described in terms of constructions that predict some grammatical and semantic features of a word class
The method extracts MWEs of different nature: collocations, colligations, constructions.

Most of the extracted MWEs are stable and frequently used, however not idiomatic.

Some part of the extracted bigrams can be described in terms of constructions that predict some grammatical and semantic features of a word class.
WHAT TO SEARCH?
This resource provides the information about how words co-occur and answers the questions like the ones above. By co-occurrences we mean:
- syntactic patterns (заниматься + Instrumentative “to be busy with”; из-за + Genitive “because of”;
  просить + Accusative/Genitive “to ask for”);
- collocations, i.e. frequently used stable expressions (тяжелая болезнь “serious illness”, печатать вслепую “to touch-type”).

HOW TO SEARCH?
Type the word (e.g., preposition до “before”) in a search box and click SEARCH. The system will show which words or grammatical features are usually used with the given word.
Be patient: it may take a few minutes to get the results.
For more information, see HELP.
<table>
<thead>
<tr>
<th>Parts of Speech</th>
<th>Mood</th>
<th>Voice</th>
<th>Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>Indicative</td>
<td>Active</td>
<td>Nominative</td>
</tr>
<tr>
<td>Adjective</td>
<td>Imperative</td>
<td>Middle</td>
<td>Vocative</td>
</tr>
<tr>
<td>Numeral</td>
<td>2nd imperative</td>
<td>Passive</td>
<td>Genitive</td>
</tr>
<tr>
<td>Numeral adjective</td>
<td>Infinitive</td>
<td></td>
<td>Genitive2</td>
</tr>
<tr>
<td>Verb</td>
<td></td>
<td></td>
<td>Dative</td>
</tr>
<tr>
<td>Participle</td>
<td></td>
<td></td>
<td>Accusative</td>
</tr>
<tr>
<td>Gerund</td>
<td></td>
<td></td>
<td>Accusative2</td>
</tr>
<tr>
<td>Adverb</td>
<td></td>
<td></td>
<td>Instrumental</td>
</tr>
<tr>
<td>Predicative</td>
<td></td>
<td></td>
<td>Locative</td>
</tr>
<tr>
<td>Parenthesis</td>
<td></td>
<td></td>
<td>Locative2</td>
</tr>
<tr>
<td>Preposition</td>
<td></td>
<td></td>
<td>Locative2</td>
</tr>
<tr>
<td>Conjunction</td>
<td></td>
<td></td>
<td>Adnomemative</td>
</tr>
<tr>
<td>Particle</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interjection</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pronoun</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjective pronoun</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adverbial pronoun</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicative pronoun</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aspect</td>
<td>Imperfective</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Perfective</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transitivity</td>
<td>Intransitive</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Transitive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tense</td>
<td>Future</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Present</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Past</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person</td>
<td>1st person</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2nd person</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3rd person</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number</td>
<td>Plural</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Singular</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>Masculine</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Feminine</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MF common</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Neuter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degree (Adj./Adverb)</td>
<td>Comparative</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Comparative2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Superlative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adj. Form</td>
<td>Full</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Short</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
на + x [Adj, Pro]+ случай

**GENDER**
- Male

**LEMMA**
- крайний: 100
- последний: 27
- всякий: 10
- этот: 6

**EXPORT**