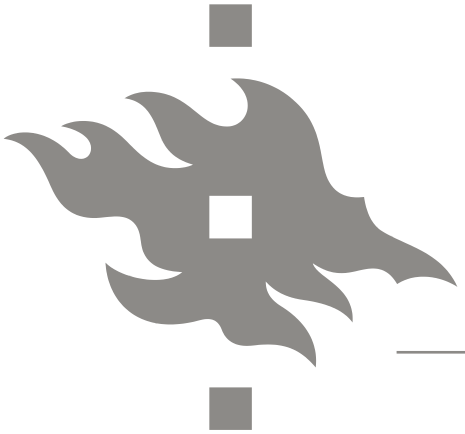


CoCoCo. automatic extraction of Russian collocations, colligations, and constructions

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**

CoCoCo

- 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?

GRET’
‘warm (up)/ heat (up)’
+ N

DUŠU ‘soul’

KROV’ ‘blood’

VODU ‘water’
MOLOKO ‘milk’
ČAJ ‘tea’

N.acc

RUKI ‘hands’
LADONI ‘palms’
NOGI ‘feet’
KOPYTA ‘hoofs’
SPINU ‘back’

MAŠINU ‘car’
MOTOR ‘motor’

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?

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’

‘to please, to make happy’

‘to warm oneself’

Collocations

Collocation typically denotes frequently repeated or statistically significant co-occurences, 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’

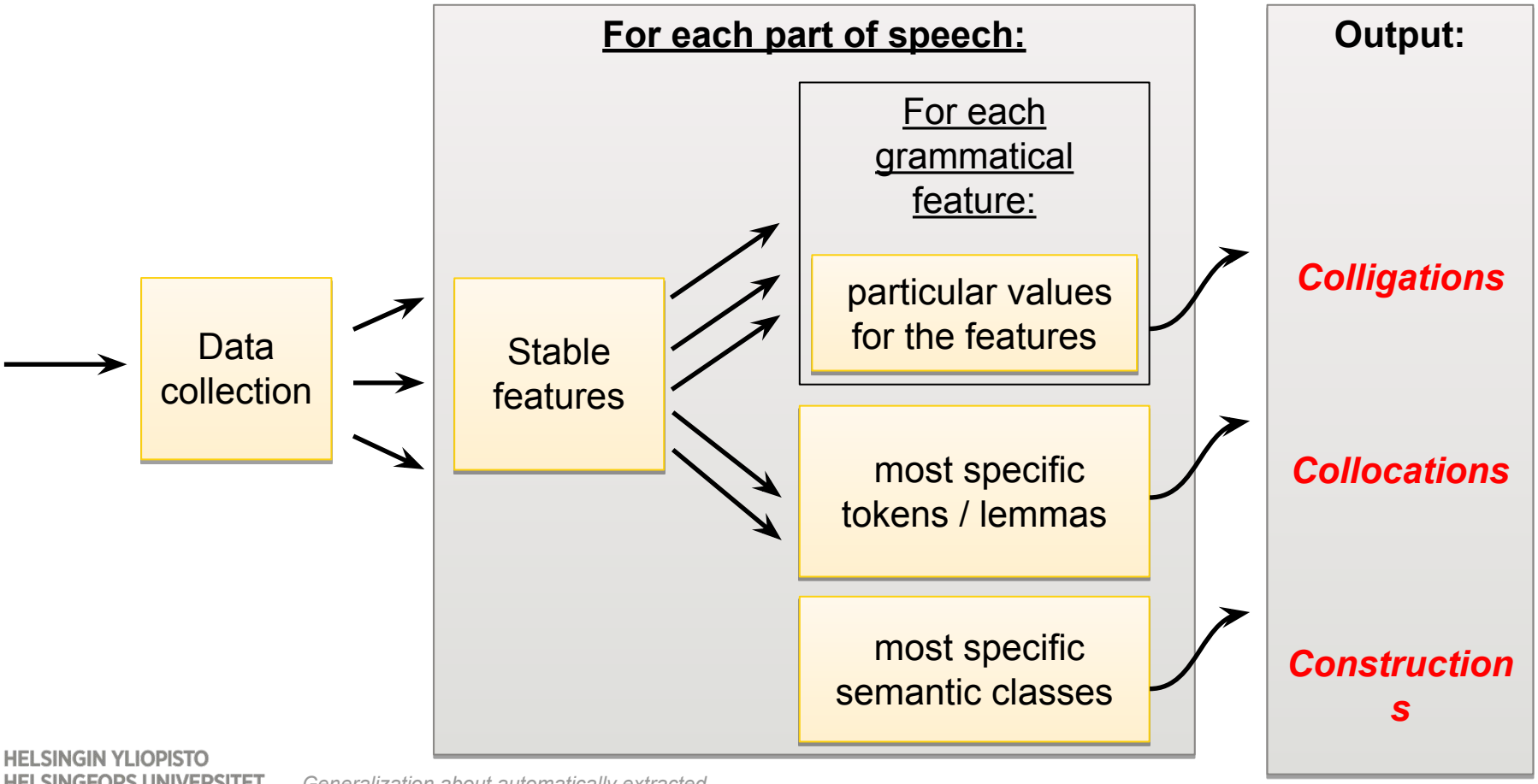
MAŠINU ‘car’
MOTOR ‘motor’

Constructions

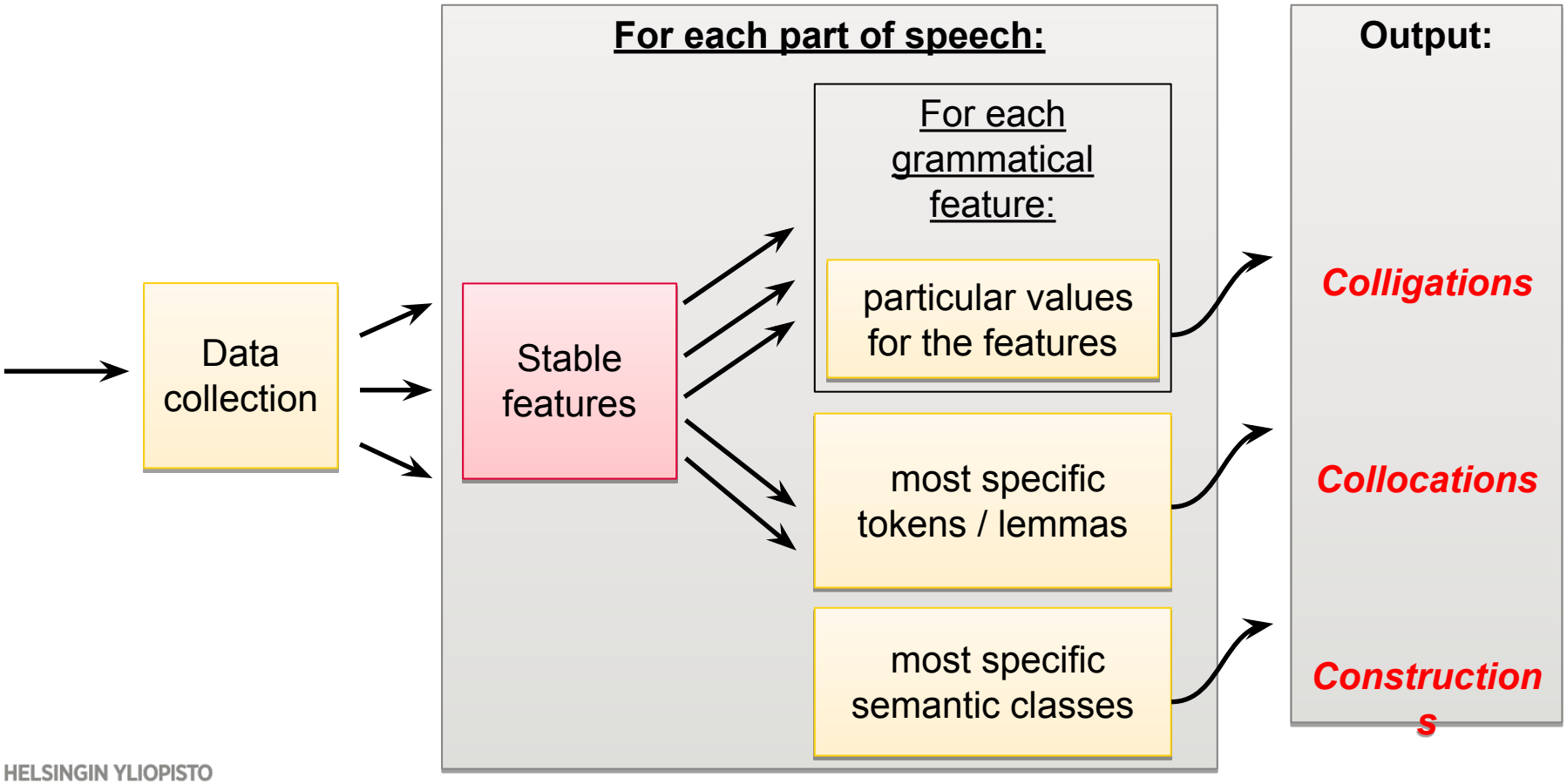
Construction – a pairing of form with meaning/use such that some aspect of the form or some aspect of the meaning/use is not strictly predictable.

(Goldberg, 1996: 68)

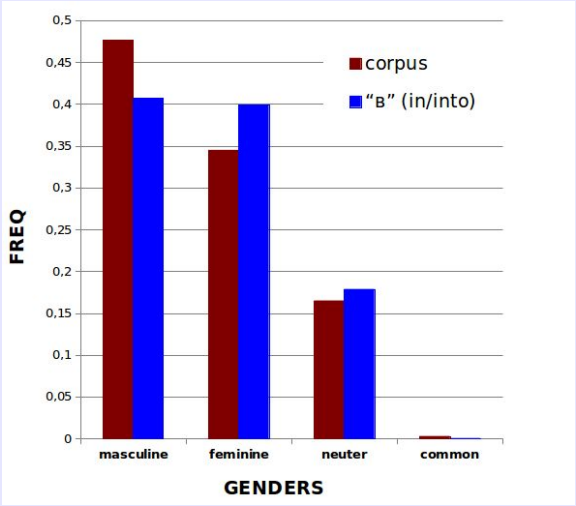
Algorithm

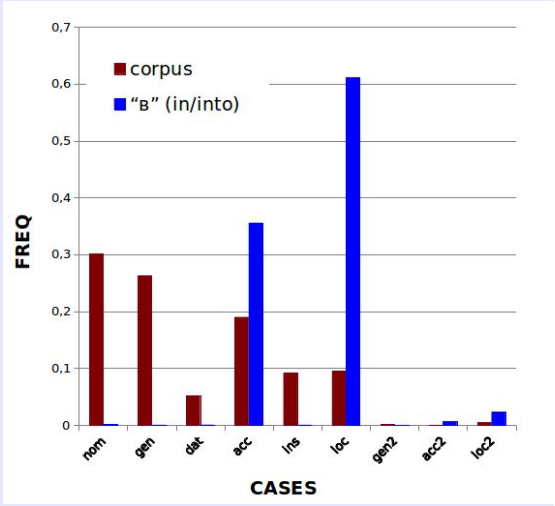

Data collection
Stable features
For each part of speech:
For each grammatical feature:
particular values for the features
most specific tokens / lemmas
most specific semantic classes
Output:
Colligations
Collocations
Constructions

Algorithm


Data collection
Stable features
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

Kullback-Leibler divergence

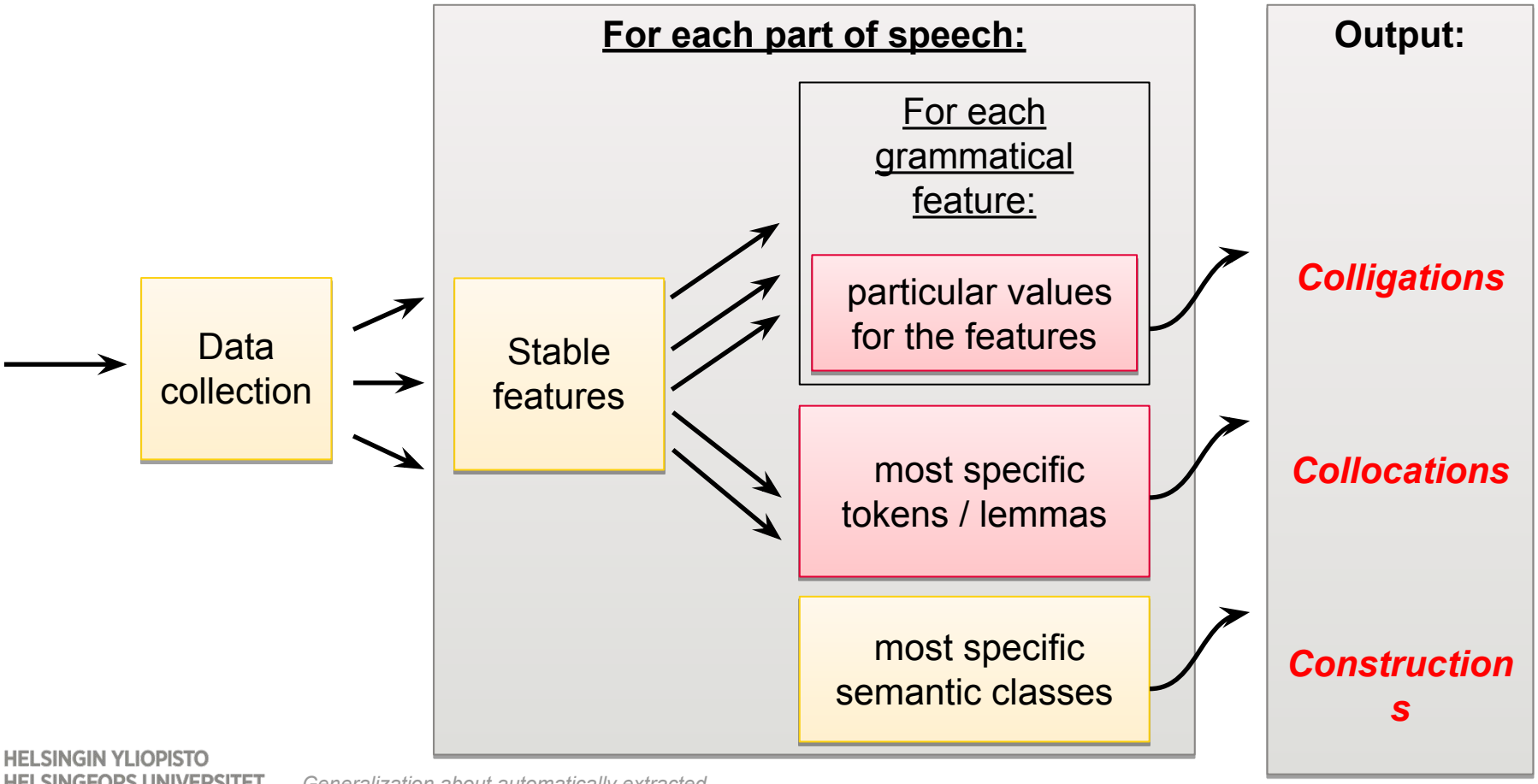




Kopotev et al. 2013

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

Algorithm


Data collection
Stable features
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

Preposition	*f*	*rFR*	*wFR*	*MI*	*Dice*	*t*
***Bez* ('Without')**	72.86	68.38	**73.34**	7.17	5.83	72.60
***U* ('Near/ At')**	3.97	1.92	**4.17**	0.00	0.00	2.92

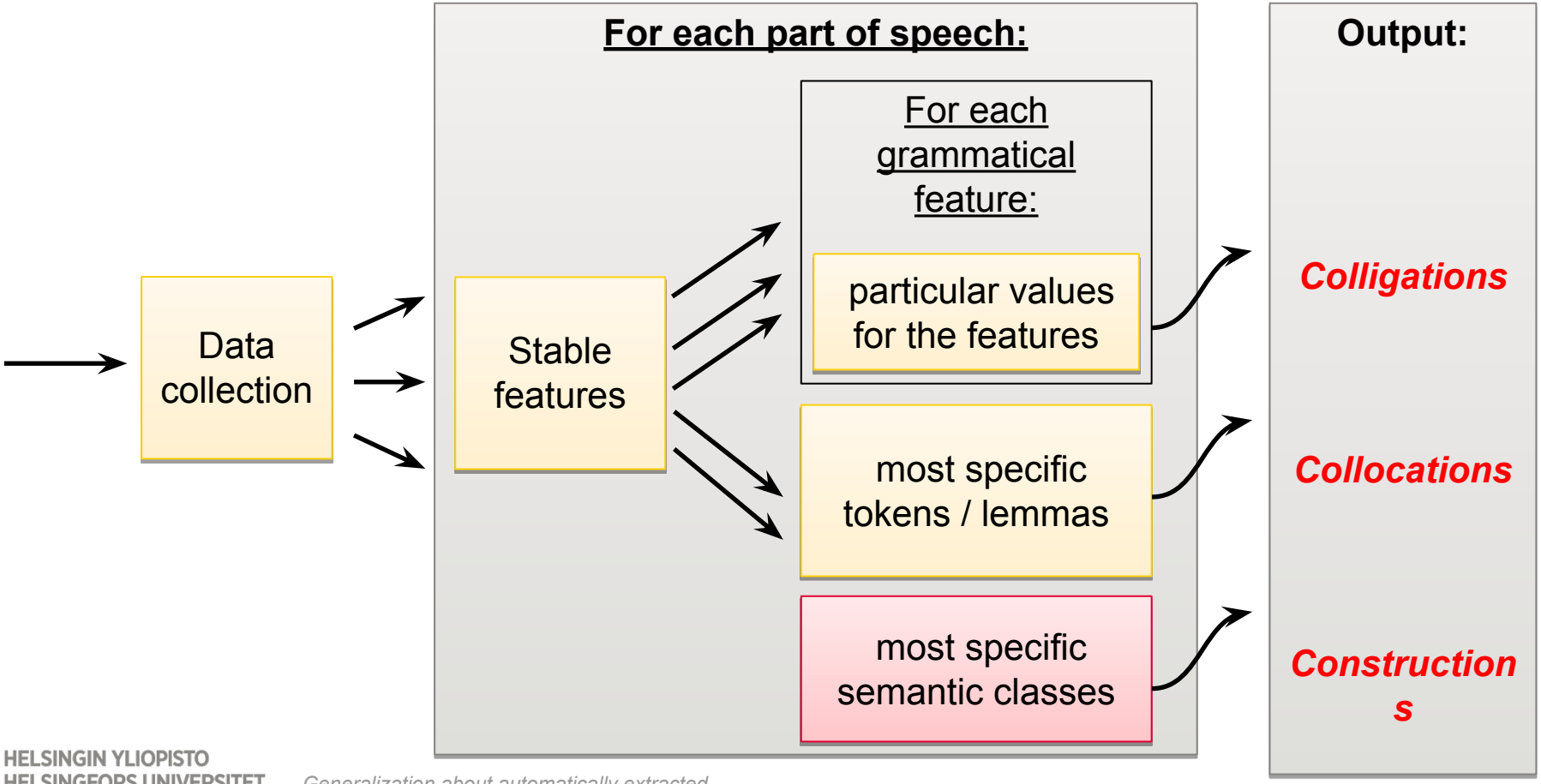
- Collocations:
  - *bez pamjati* (without.PREP memory.NOUN.SG.GEN, 'like mad', 'passionately')
  - *bez ceremonij* (without.PREP ceremony.NOUN.PL.GEN, 'informally')
  - *u istokov* (at.PREP river source.NOUN.PL.GEN, 'at the origins')

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'co* 'porch', *stena* 'wall', etc.;
  - 13: [*u* 'near/at/-' + ANIMAL]: *koška* 'cat', *korova* 'cow', *mlekopitajuščee* 'mammal', etc.;
  - 10: [*u* 'near/at/-' + RELATIVE]: *rebenok* 'child', *papa* 'dad', *tešča* 'mother in low', 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


Data collection
Stable features
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

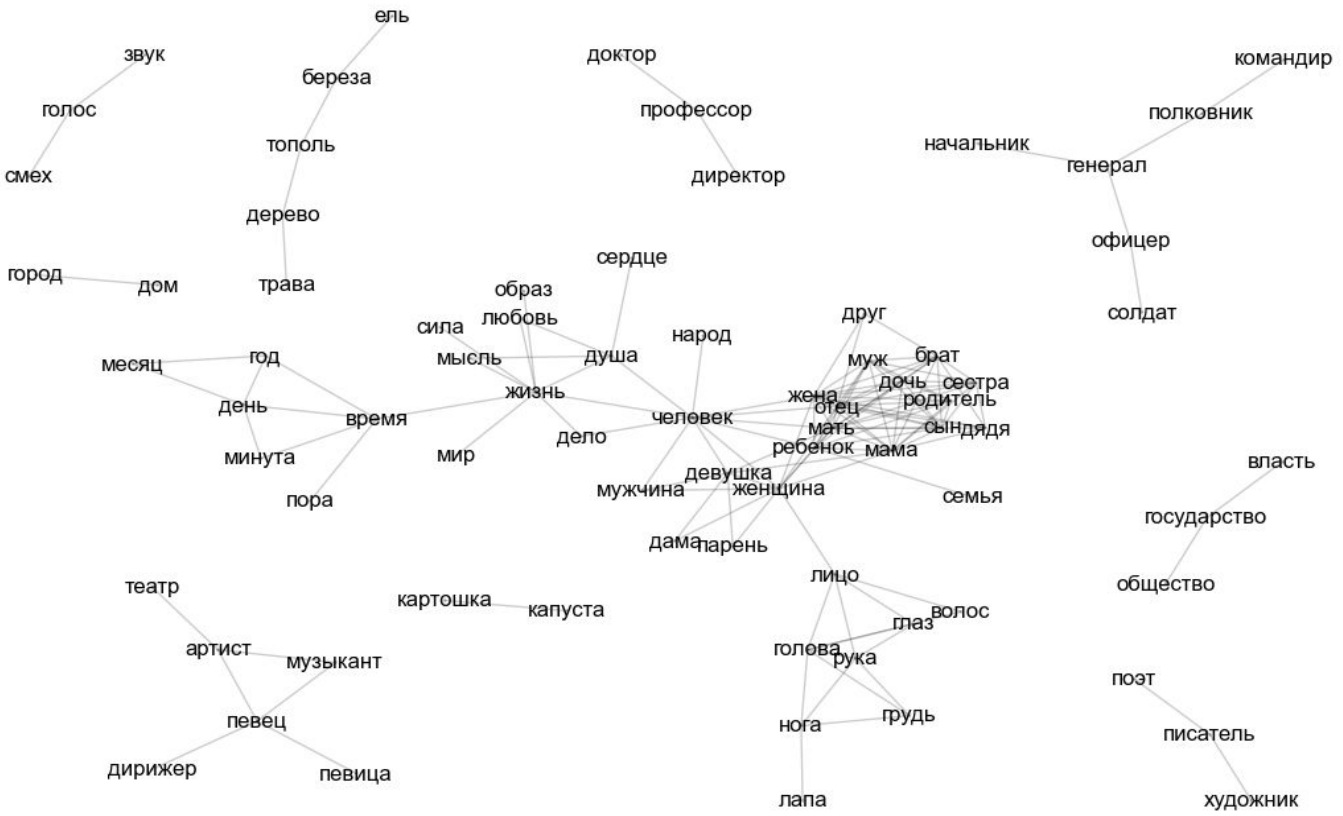
Semantic clustering method

"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*.

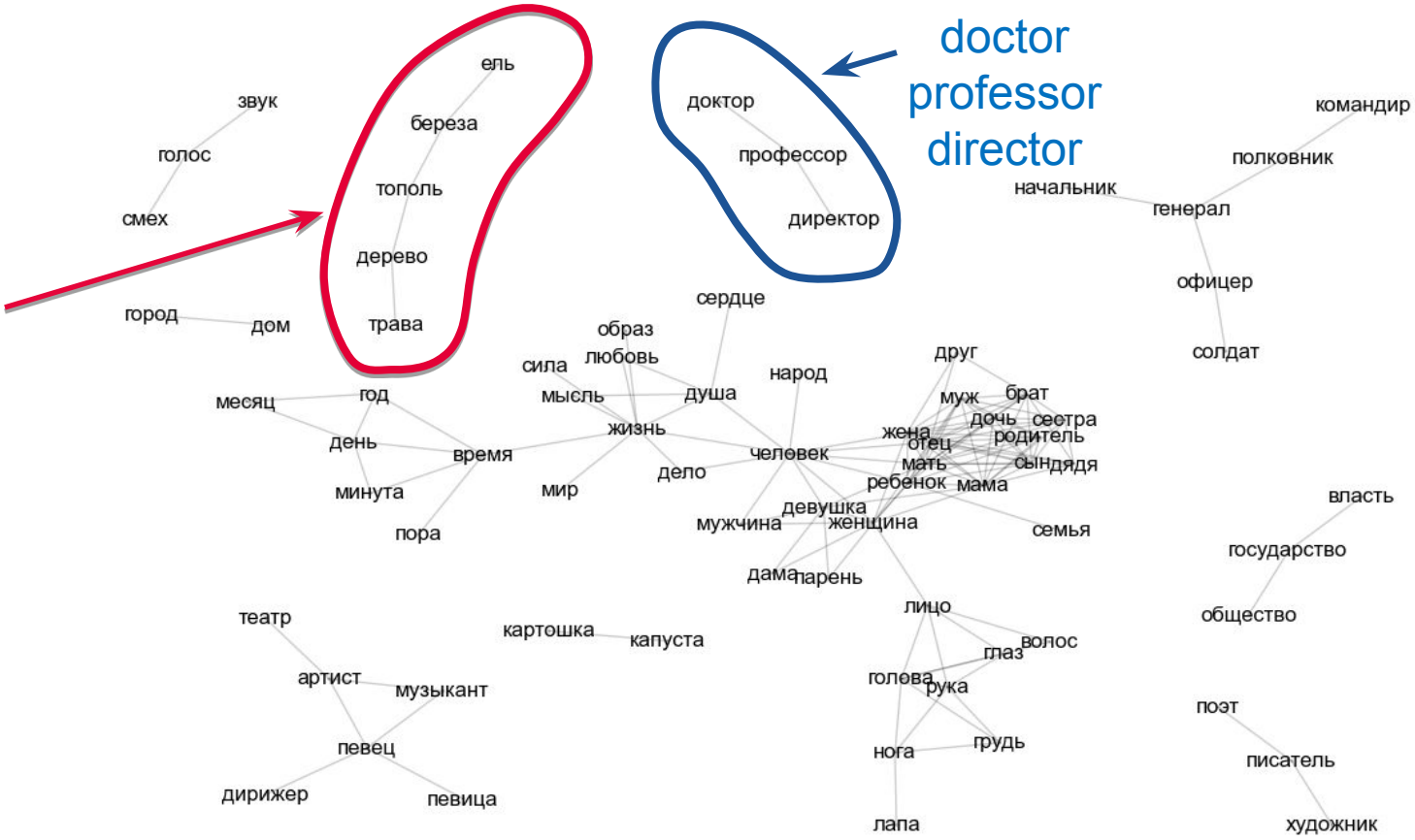
Constructional profile for [*molodoj* ‘young’ + X]

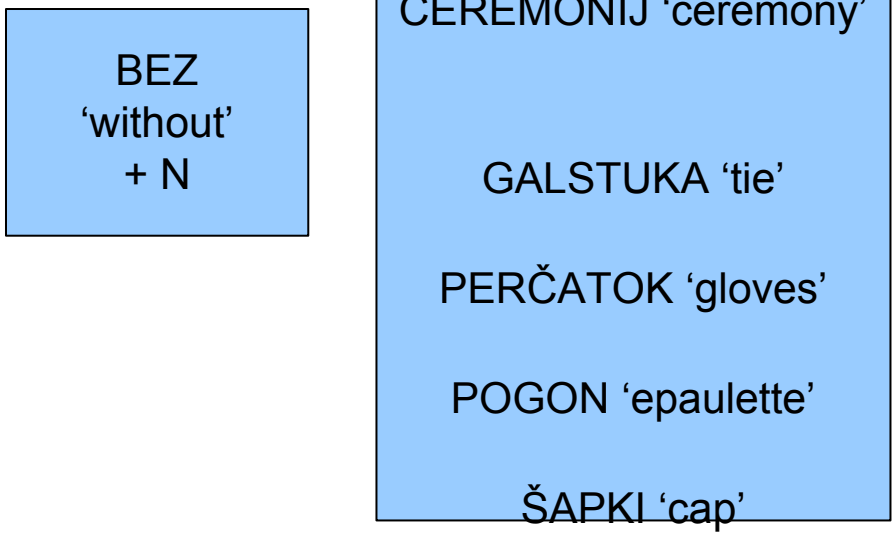


Constructional profile for [*molodoj* ‘young’ + X]

spruce
birch
poplar
tree
grass

doctor
professor
director




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

BEZ
‘without’
+ N

PAMJATI ‘memory’

CEREMONIJ ‘ceremony’

GALSTUKA ‘tie’

PERČATOK ‘gloves’

POGON ‘epaulette’

ŠAPKI ‘cap’

BEZ
‘without’
+ N

PAMJATI ‘memory’

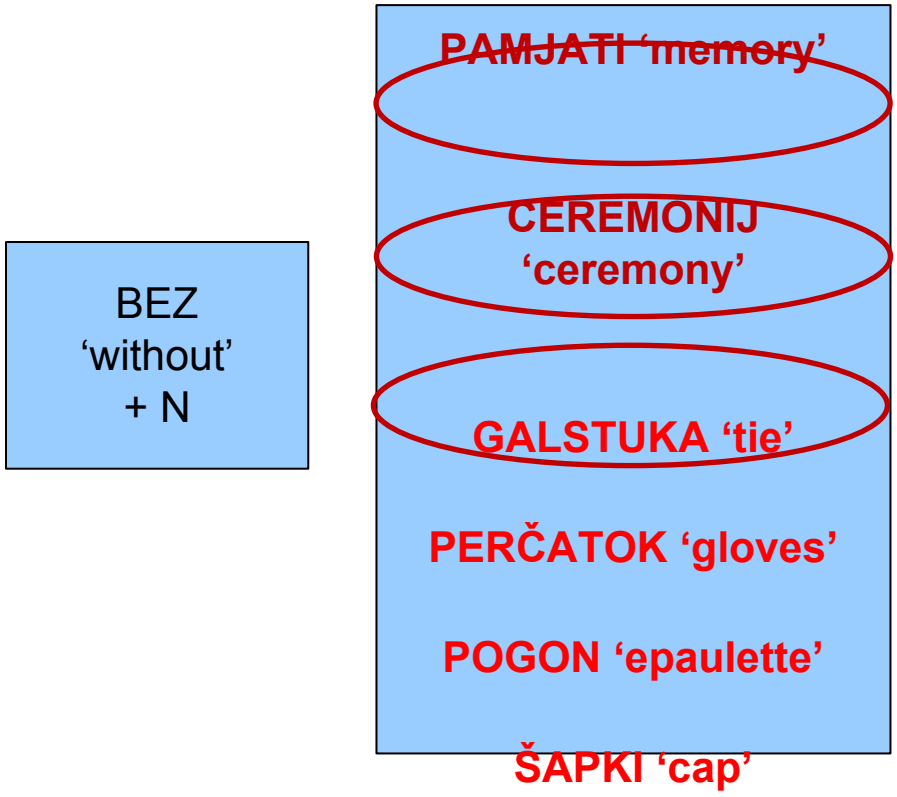
CEREMONIJ
‘ceremony’

GALSTUKA ‘tie’

PERČATOK ‘gloves’

POGON ‘epaulette’

ŠAPKI ‘cap’


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

RNC
I-RU

Enter a word

OPTIONS

Enter a word

OPTIONS

+

SEARCH CLEAR

Русские просят совет или совета?
Что по-русски делают вслепую?

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.

PARTS OF SPEECH

- [ ] Noun
- [ ] Adjective
- [ ] Numeral
- [ ] Numeral adjective
- [x] Verb
- [ ] Participle
- [ ] Gerund
- [ ] Adverb
- [ ] Predicative
- [ ] Parenthesis
- [ ] Preposition
- [ ] Conjunction
- [ ] Particle
- [ ] Interjection
- [ ] Pronoun
- [ ] Adjective pronoun
- [ ] Adverbial pronoun
- [ ] Predicative pronoun

MOOD

- [ ] Indicative
- [ ] Imperative
- [ ] 2nd imperative
- [ ] Infinitive

ASPECT

- [ ] Imperfective
- [ ] Perfective

TRANSITIVITY

- [ ] Intransitive
- [ ] Transitive

TENSE

- [ ] Future
- [ ] Present
- [ ] Past

PERSON

- [ ] 1st person
- [ ] 2nd person
- [ ] 3rd person

VOICE

- [ ] Active
- [ ] Middle
- [ ] Passive

NUMBER

- [ ] Plural
- [ ] Singular

GENDER

- [ ] Masculine
- [ ] Feminine
- [ ] MF common
- [ ] Neuter

DEGREE (ADJ./ADVERB)

- [ ] Comparative
- [ ] Comparative2
- [ ] Positive
- [ ] Superlative

ADJ. FORM

- [ ] Full
- [ ] Short

CASE

- [ ] Nominative
- [ ] Vocative
- [ ] Genitive
- [ ] Genitive2
- [ ] Dative
- [ ] Accusative
- [ ] Accusative2
- [ ] Instrumental
- [ ] Locative
- [ ] Locative2
- [ ] Adnumerative

ANIMACY

- [ ] Animate
- [ ] Inanimate

OK CLEAR ?

RNC
I-RU

на | Choose options | случай

OPTIONS | OPTIONS | OPTIONS

SEARCH CLEAR

на + x [Adj, Pro]+ случай

GENDER

Male

LEMMA

крайний	100
последний	27
всякий	10
этот	6

EXPORT