Crossing Corpora. Modelling Semantic Similarity across Languages and Lects.

Yves Peirsman
Supervisors: Dirk Geeraerts & Dirk Speelman

Quantitative Lexicology and Variational Linguistics

Background

- Corpus-based investigation of lexical variation

JEANS	jeans	jeansbroek	spijkerbroek
BE	35%	45%	20%
NL	20%	15%	65%

- Problem: semantic equivalence
- Solution: distributional models of semantic similarity
- Extension from one corpus to two corpora
- Application to two corpora from the same language: variational linguistics
- Application to two corpora from different languages: cross-language knowledge induction

Outline

Outline

Distributional Models

Distributional hypothesis

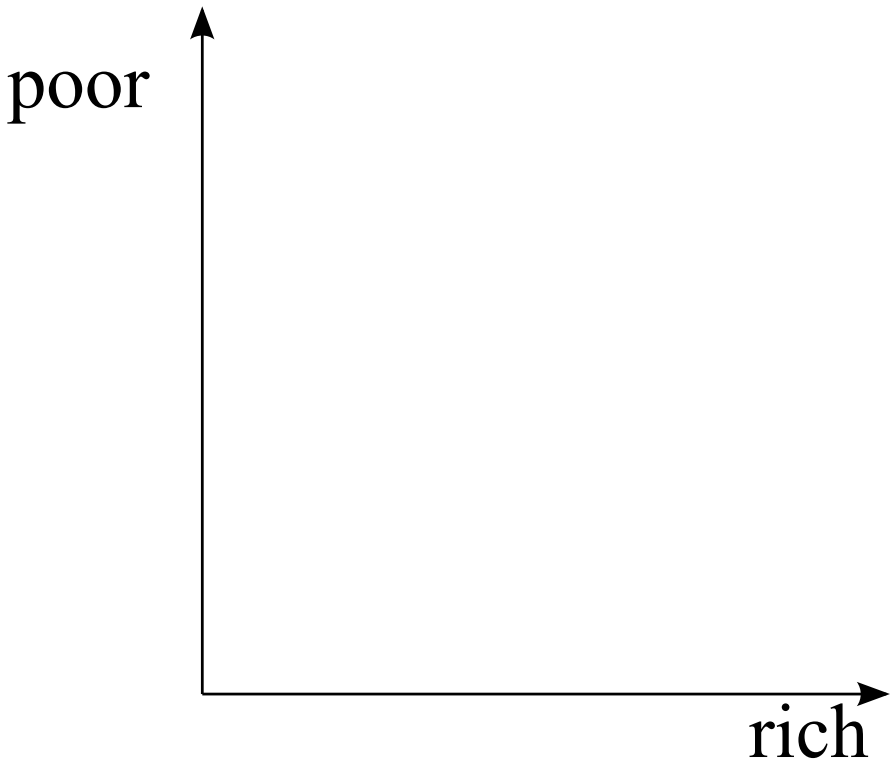
Semantically similar words appear in similar contexts
(Wittgenstein, Firth, Harris)

Recipe

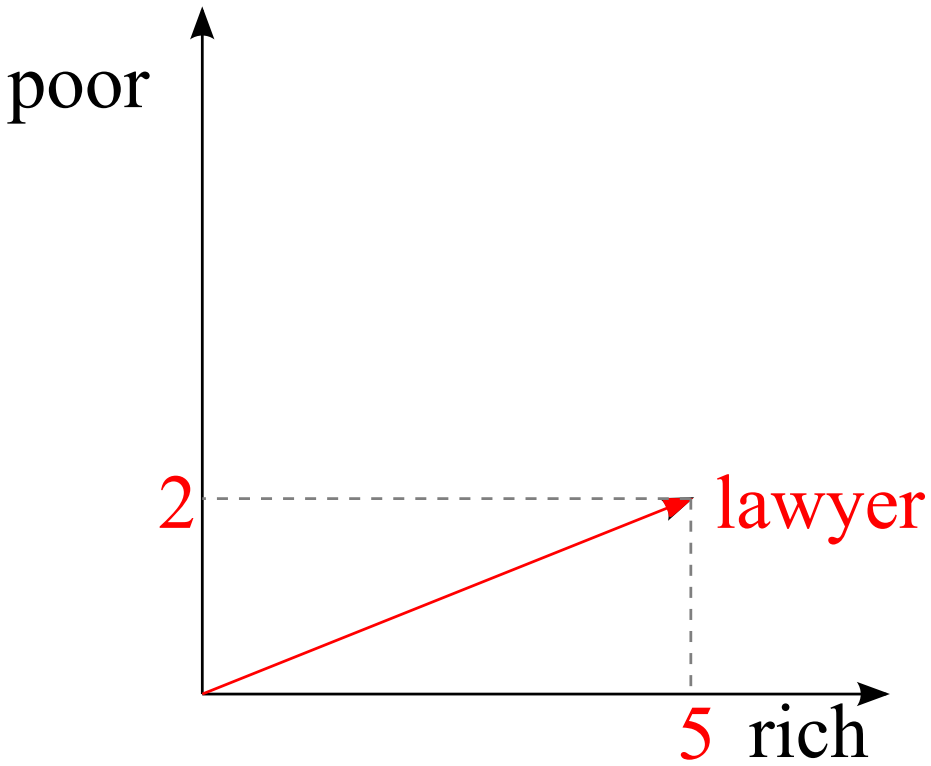
Ingredients: 1 large corpus, 1 definition of context, target words

- search the corpus for the contexts of your target words
- calculate the contextual similarity between each pair of target words
- find the most contextually similar target words

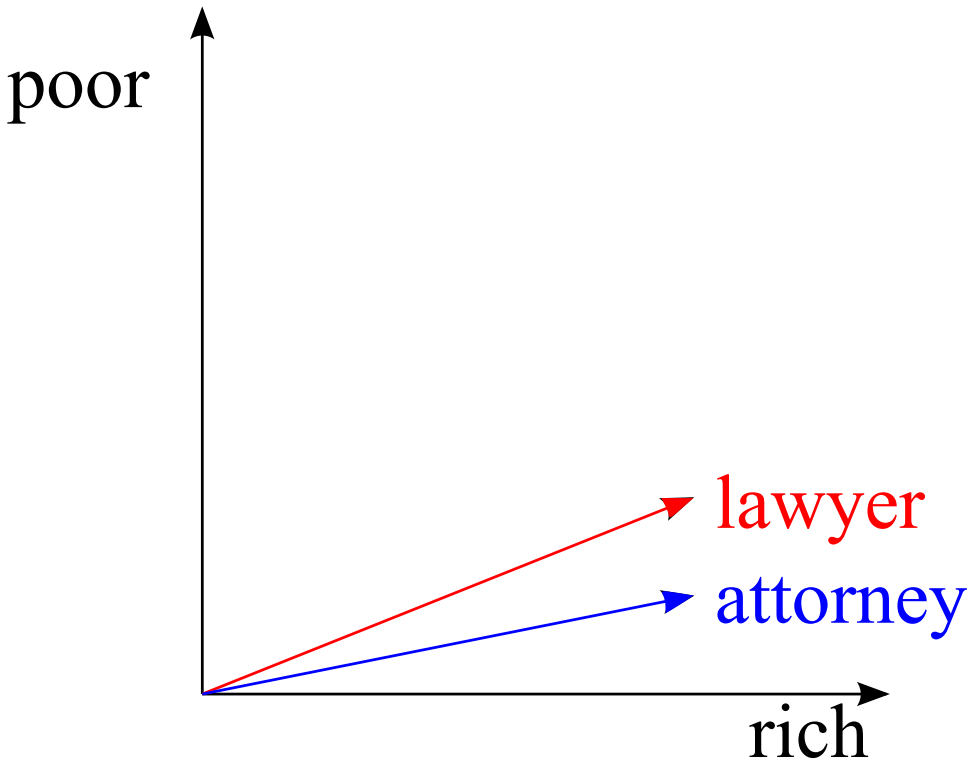
Distributional Models



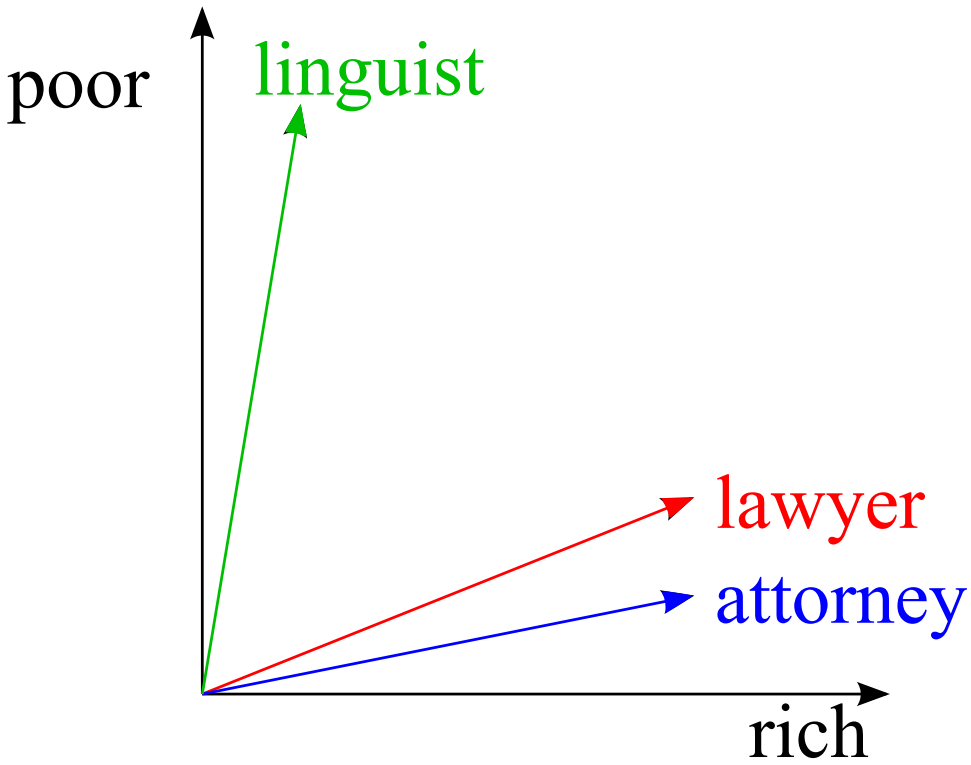
Distributional Models



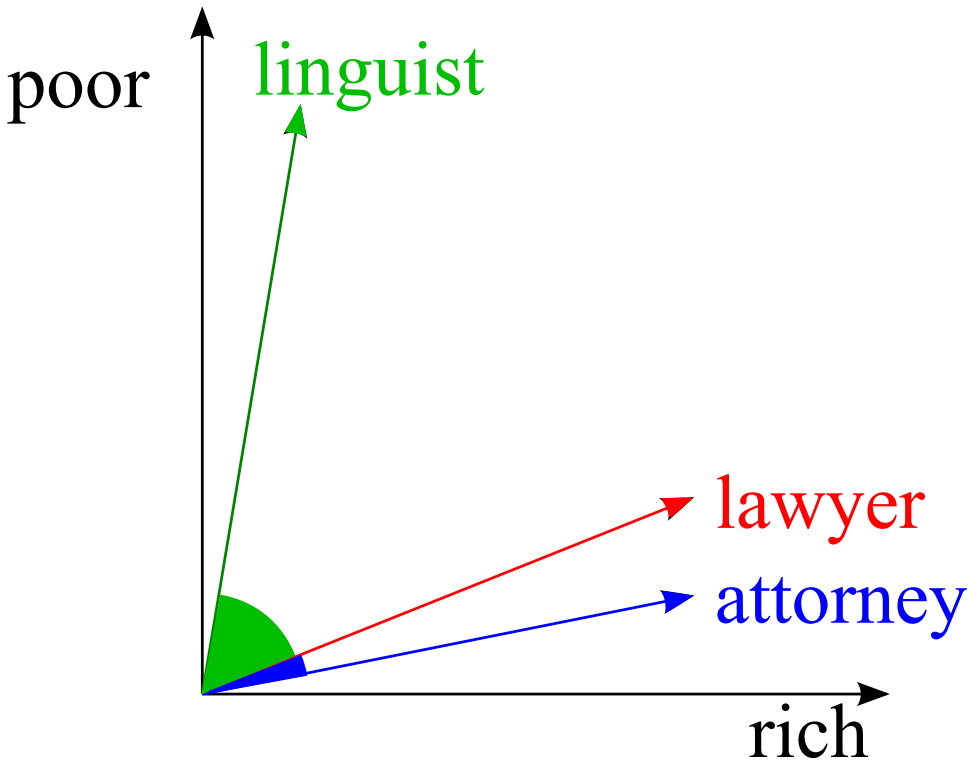
Distributional Models



Distributional Models



Distributional Models



Distributional Models

Definition of context

- context words: rich, poor
  $\longrightarrow$ several context sizes
- syntactic relations: subject of plead, object of eat
- documents or paragraphs

The definition of context influences the type of semantic relation that is modelled.

3 evaluation exercises

- human similarity judgements
- EuroWordNet
- free associations

Distributional Models

Human similarity judgements

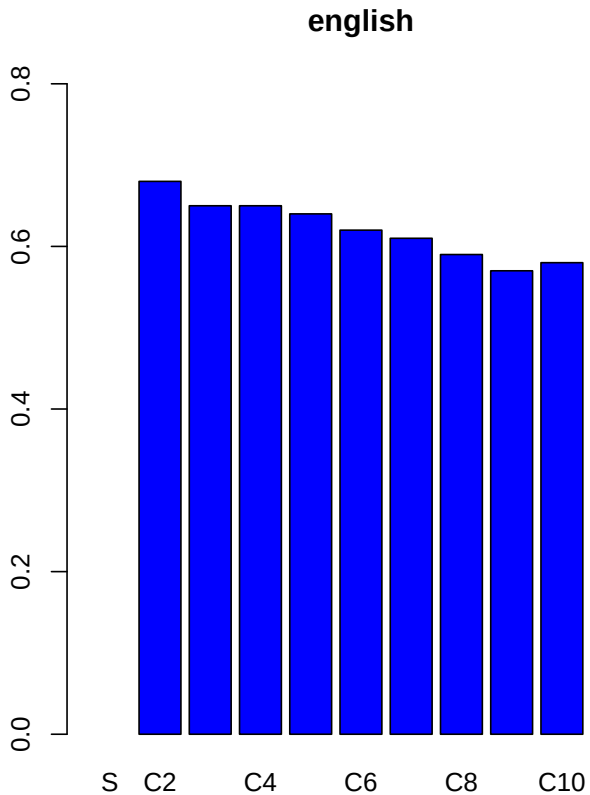
autograph	shore	0.06	piloot	fornuis	1.48
monk	slave	0.57	sla	reiger	2.28
glass	jewel	1.78	konijn	vlieg	8.79
bird	crane	2.63	grasmachine	stofzuiger	10.59
cemetery	graveyard	3.88	goudvis	haring	14.45
gem	jewel	3.94	clementine	sinaasappel	17.90

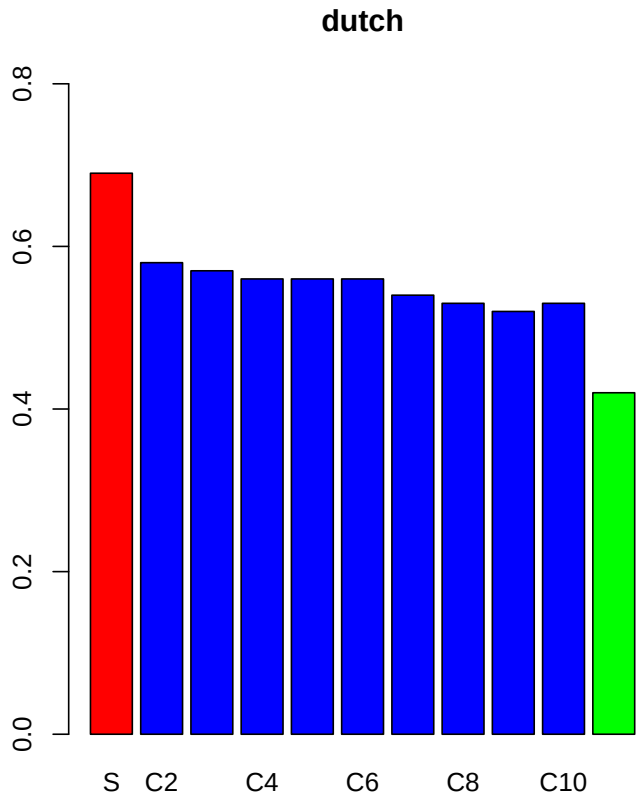
Data

- Dutch: 250M words of Belgian Dutch newspaper language
- English: BNC

Distributional Models

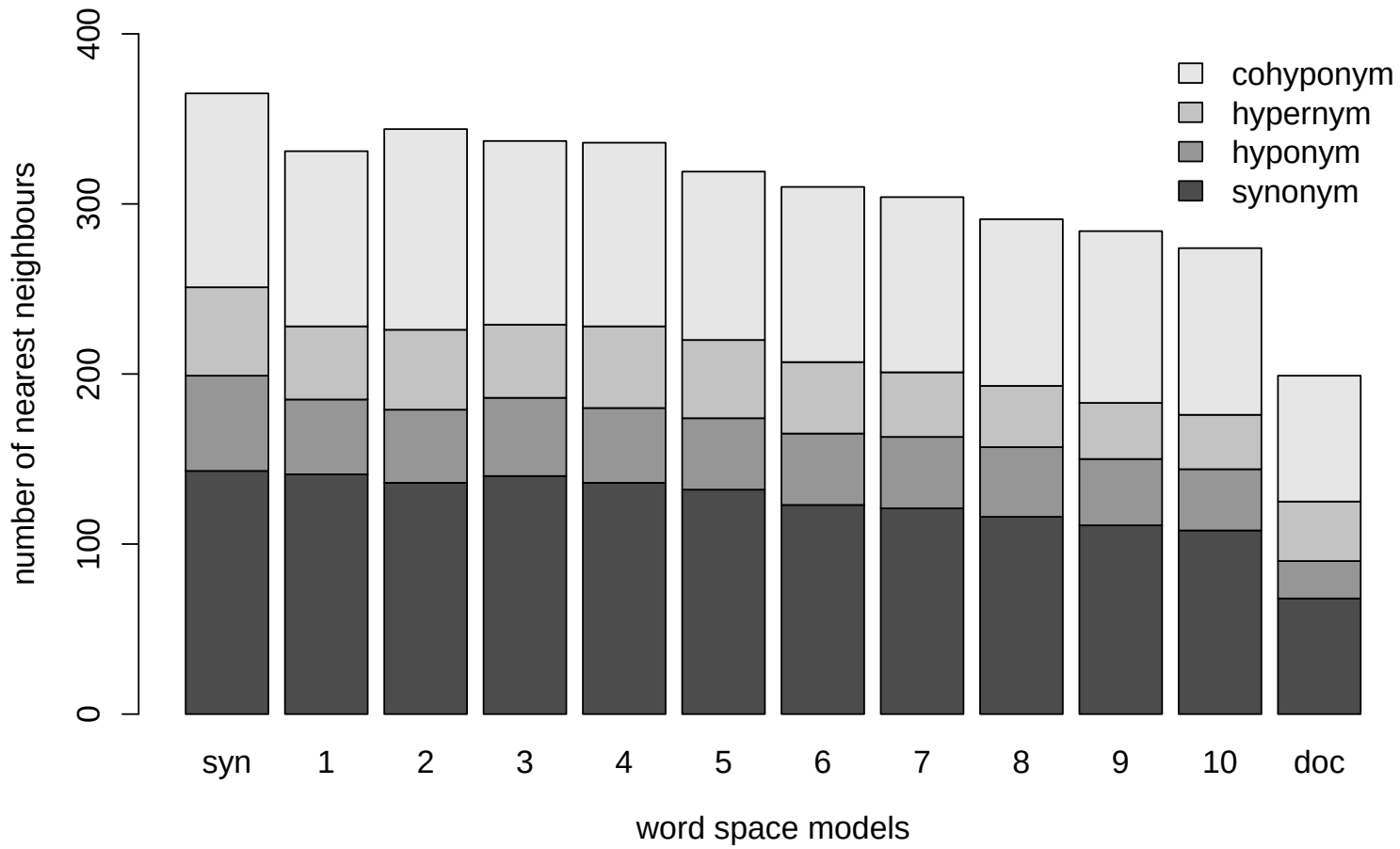
Human similarity judgements





Distributional Models

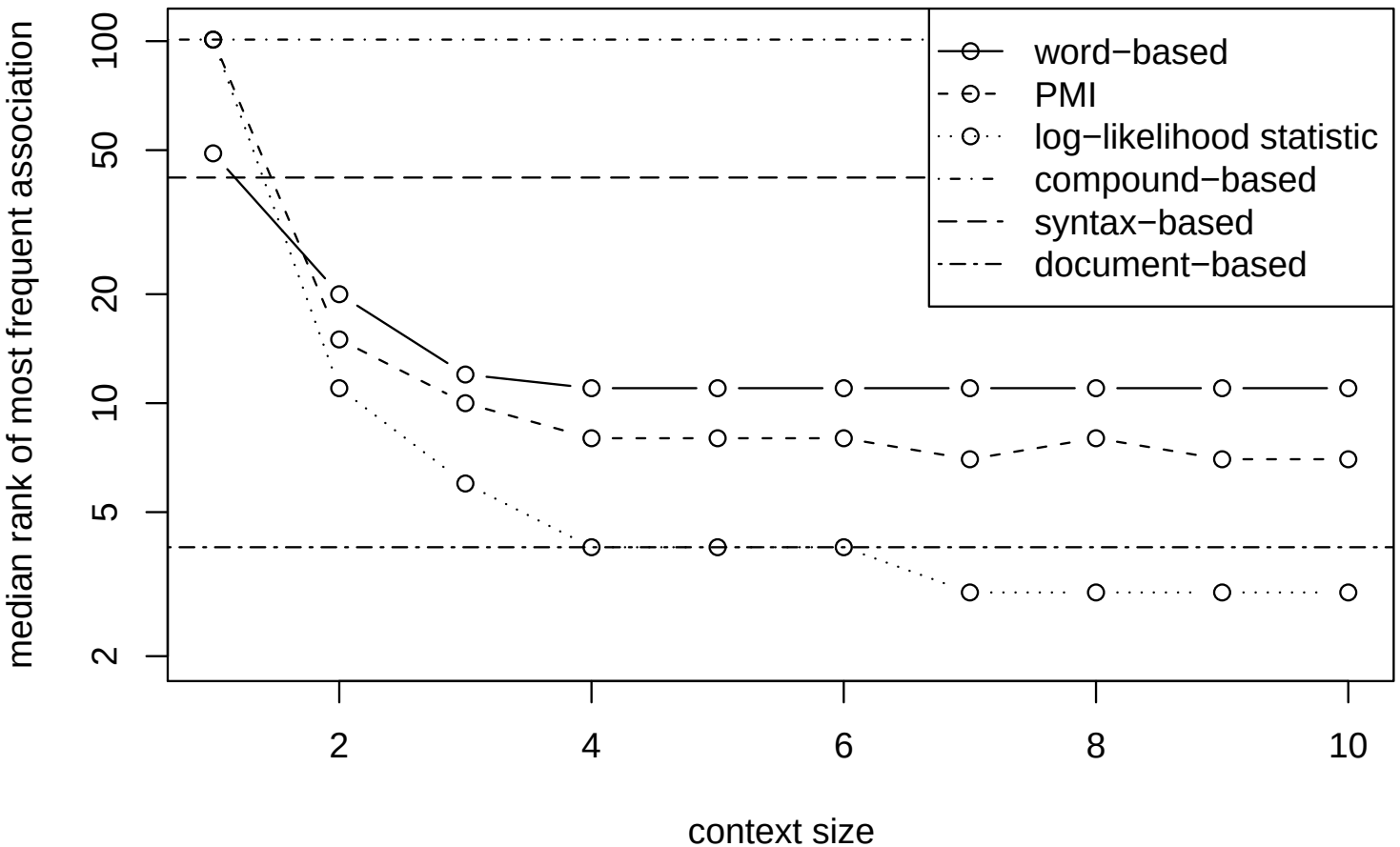
EuroWordNet



Distributional Models

Free associations

strawberry – red, wave – sea, pepper – salt



Distributional Models

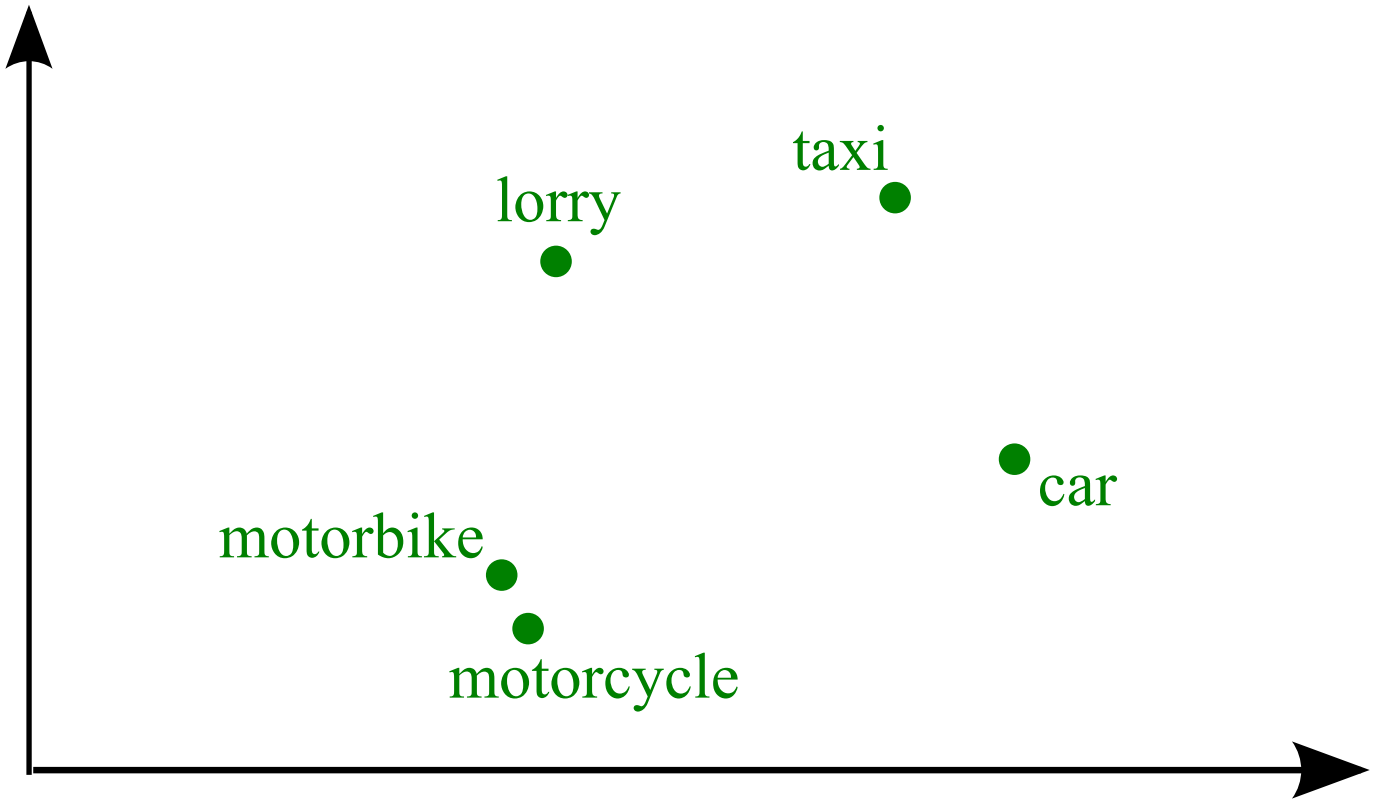
Conclusions

- Semantic similarity: syntax-based models and word-based models with small context sizes
- Semantic relatedness: document-based models
- Problem of data sparseness

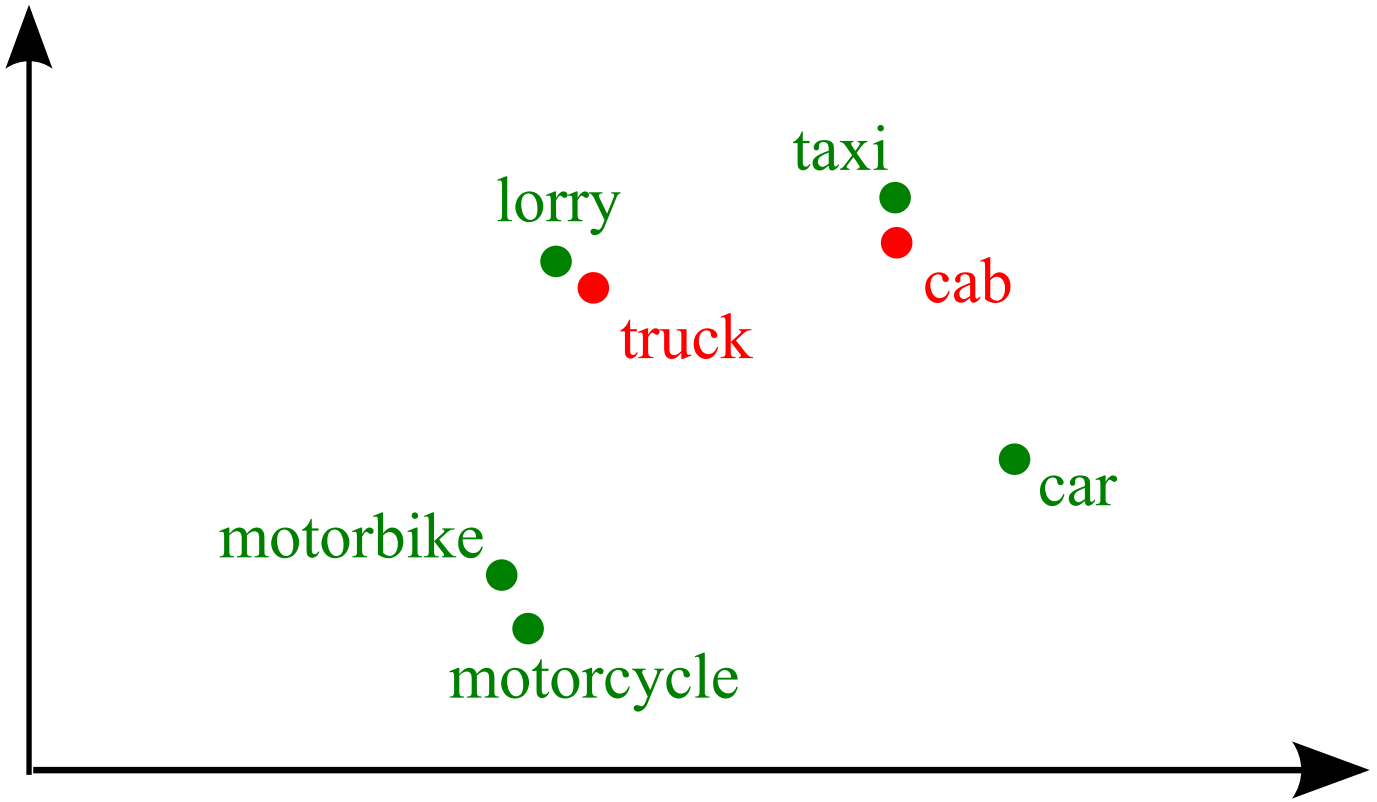
Modelling semantic similarity across languages and lects: word-based models with small context sizes

Outline

Cross-lectal similarity



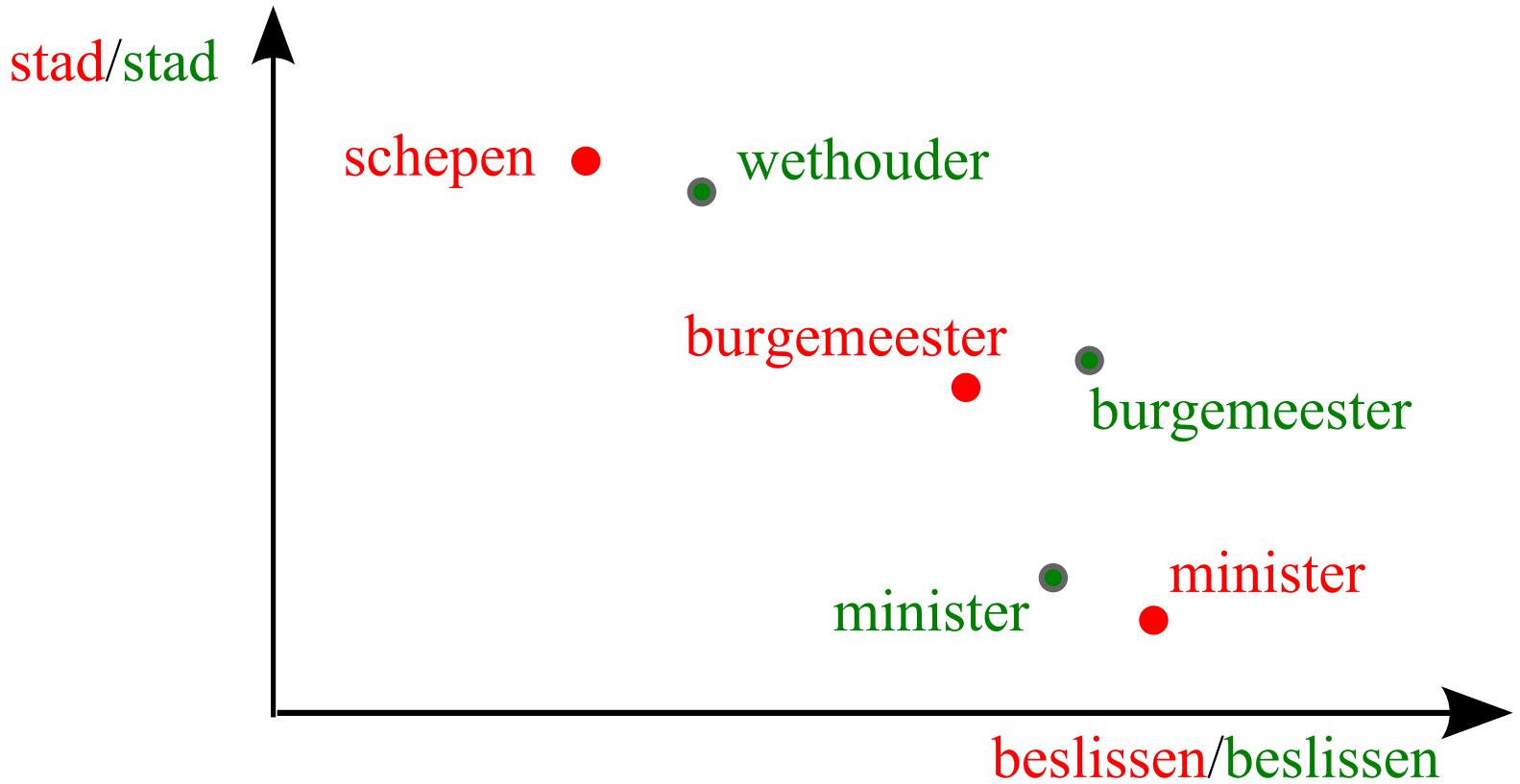
Cross-lectal similarity



Cross-lectal similarity

- Extension of semantic similarity to two corpora
- Requirement: context features are identical between the corpora
- Bilectal models can help us
  - recognize markers of a specific language variety
  - extract the synonym from another language variety

Cross-lectal similarity



Cross-lectal similarity

Lectal markers

- traditional frequency-based definition of keywords
  - different frequencies ⇒ lectal markers?
  - lectal markers ⇒ different frequencies?
- new distributional definition

Data

250M words of newspaper language from BE and NL

Referentiebestand Belgisch Nederlands

list of 4,000 markers of Belgian Dutch with their Netherlandic Dutch alternative

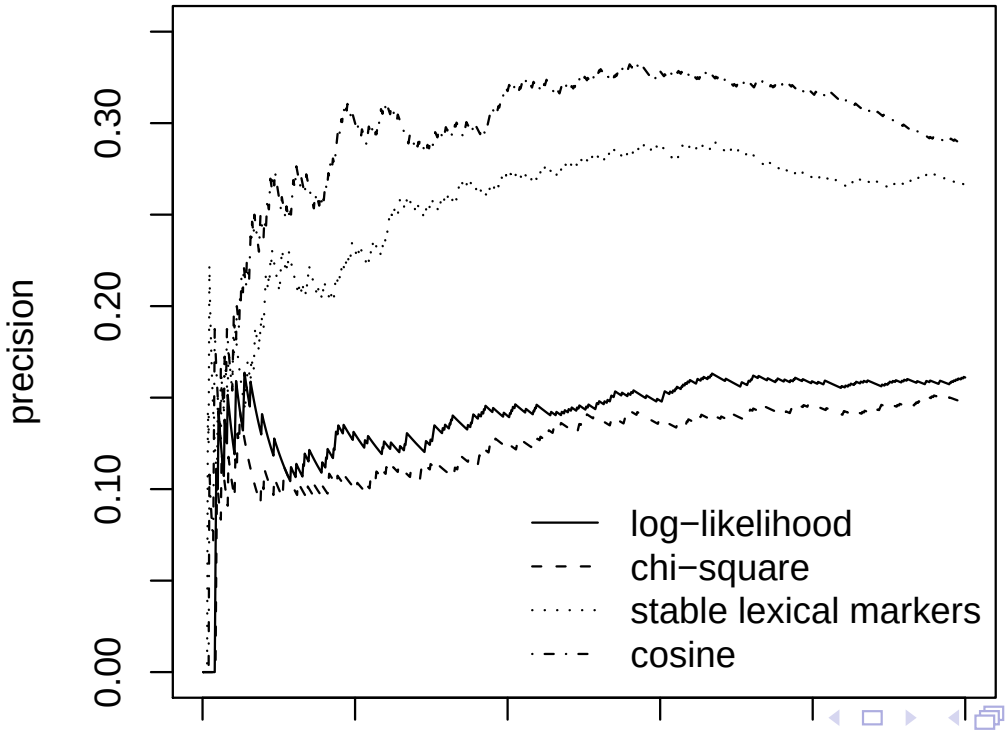
Cross-lectal similarity

alternation	BE	NL	meaning
free	ajuin	ui	'onion'
	aftrekker	flesopener	'bottle opener'
unique	doctorandus	promovendus	'PhD student'
	schepen	wethouder	'alderman'
unlexicalised	dodentol	—	'number of victims'
	vieruurtje	—	'afternoon snack'
informal	refter	eetzaal	'canteen'
	schoon	mooi	'beautiful'
restrictions	schacht	—	'first-year student'
	kaka	poep	'poop'
cultural	CM	—	Belgian health fund
	rijkswacht	—	former national police

Cross-lectal similarity

Lectal markers

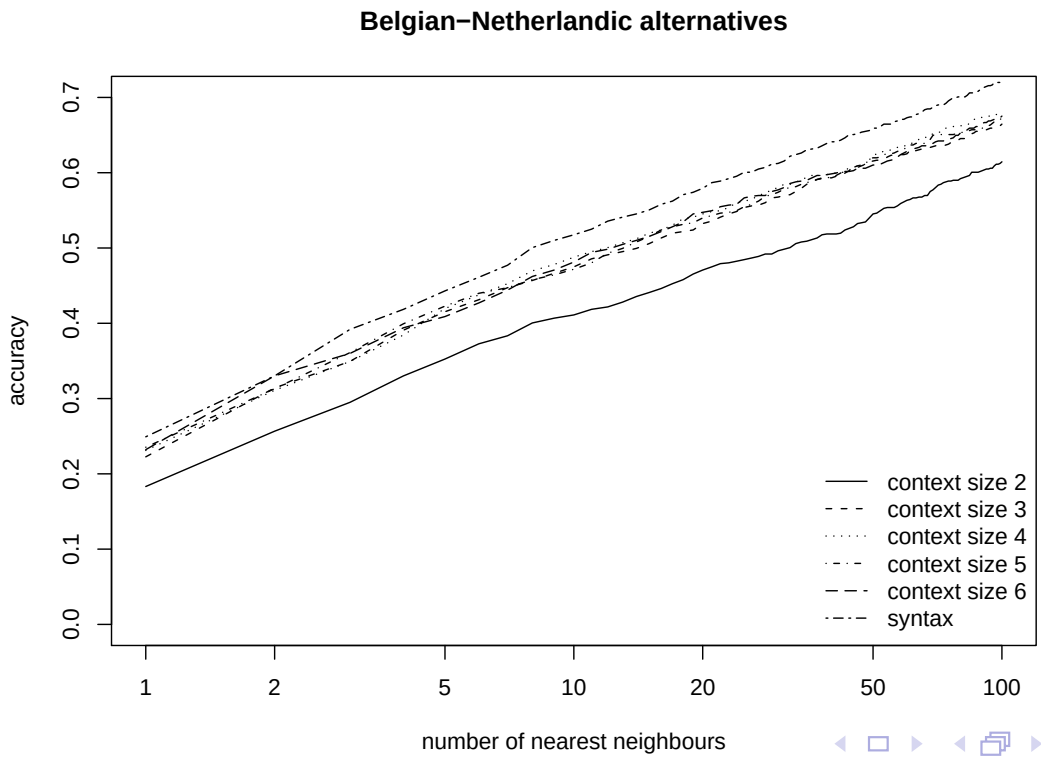
- Look for words with an unexpectedly low cross-lectal distributional similarity
- Method finds more RBBN words than traditional keyword methods



Cross-lectal similarity

Cross-lectal synonyms

- Extract nearest Netherlandic neighbours to RBBN words.
- $>$ 20% of the 1st nearest neighbours are RBBN synonyms.
- $>$ 50% of the RBBN synonyms are in the top 15 neighbours.



Cross-lectal similarity

Examples of correct pairs

Belgian marker	Netherlandic alternative	meaning
bankbriefje	bankbiljet	'bank note'
confituur	jam	'jam'
fier	trots	'proud'
gamma	assortiment	'assortment'
job	baan	'job'
kot	studentenkamer	'student room'
living	woonkamer	'living room'
microgolf	magnetron	'microwave oven'
proper	schoon	'clean'
uitbaten	exploiteren	'exploit'

Cross-lectal similarity

Problem group 1: polysemous words

Belgian marker	alternative	nearest neighbours
katholiek ‘catholic’	goed ‘good’	katholiek ‘catholic’ rooms-katholiek ‘roman catholic’ protestants ‘protestant’
stelling ‘position’	steiger ‘scaffold’	stelling ‘position’ standpunt ‘position’ opvatting ‘opinion’
tenor ‘tenor’	topper ‘steersman’	tenor ‘tenor’ countertenor ‘counter tenor’ bariton ‘baritone’

Cross-lectal similarity

Problem group 2: colloquial words

Belgian marker	alternative	nearest neighbours
bal ‘ball/franc’	frank ‘franc’	bal ‘ball’ schot ‘shot’ rust ‘rest’
flessen ‘fail’	zakken	aanlengen ‘dilute’ hergebruiken ‘reuse’ recycle ‘recycle’
boerenbuiten ‘countryside’	platteland	vrouwenhuis ‘women’s house’ buitenlieden ‘countrymen’ huisschrijver ‘resident author’

Cross-lectal similarity

Discussion

- Distributional approaches can model lexical variation
- Tool for variational-linguistic research
- Polysemy is a major problem

Future work

- Application to other languages, other genres
- Addressing polysemy

Outline

Cross-lingual similarity

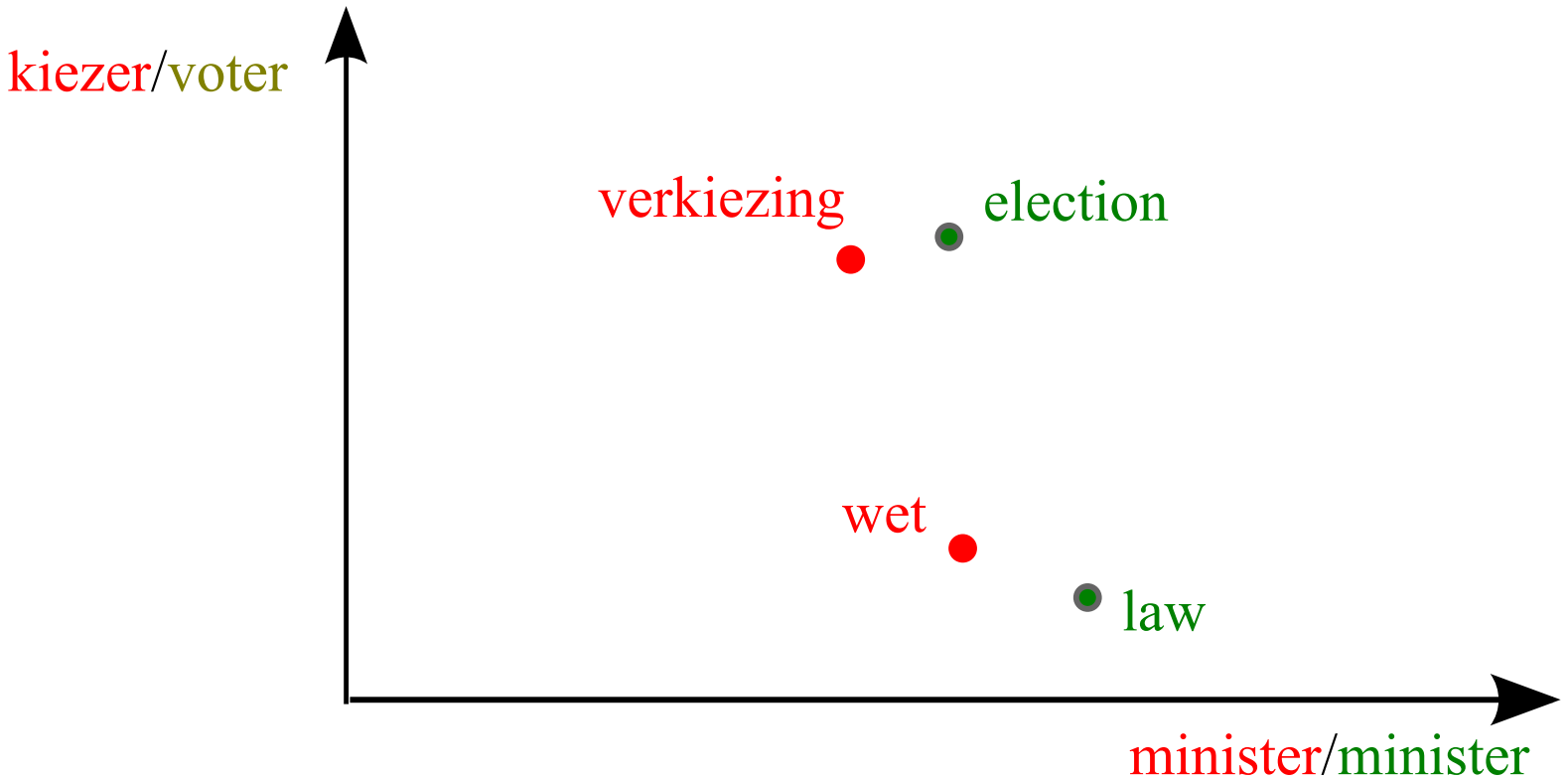
- Application of distributional models to two corpora can be extended to two languages
- Condition of overlap in context features
- Many languages share a number of words (cognates, loans)
- These words generally have a similar meaning.

Cross-lingual similarity

Method

1. Identify the shared words between the two languages.
   e.g. manager, school, kind
2. Use these words as corresponding context words to find a small number of reliable translations
   e.g. auto–car, boek–book, kind–child
3. Add these newly found translations to the set of corresponding context words
4. Repeat

Cross-lingual similarity



Cross-lingual similarity

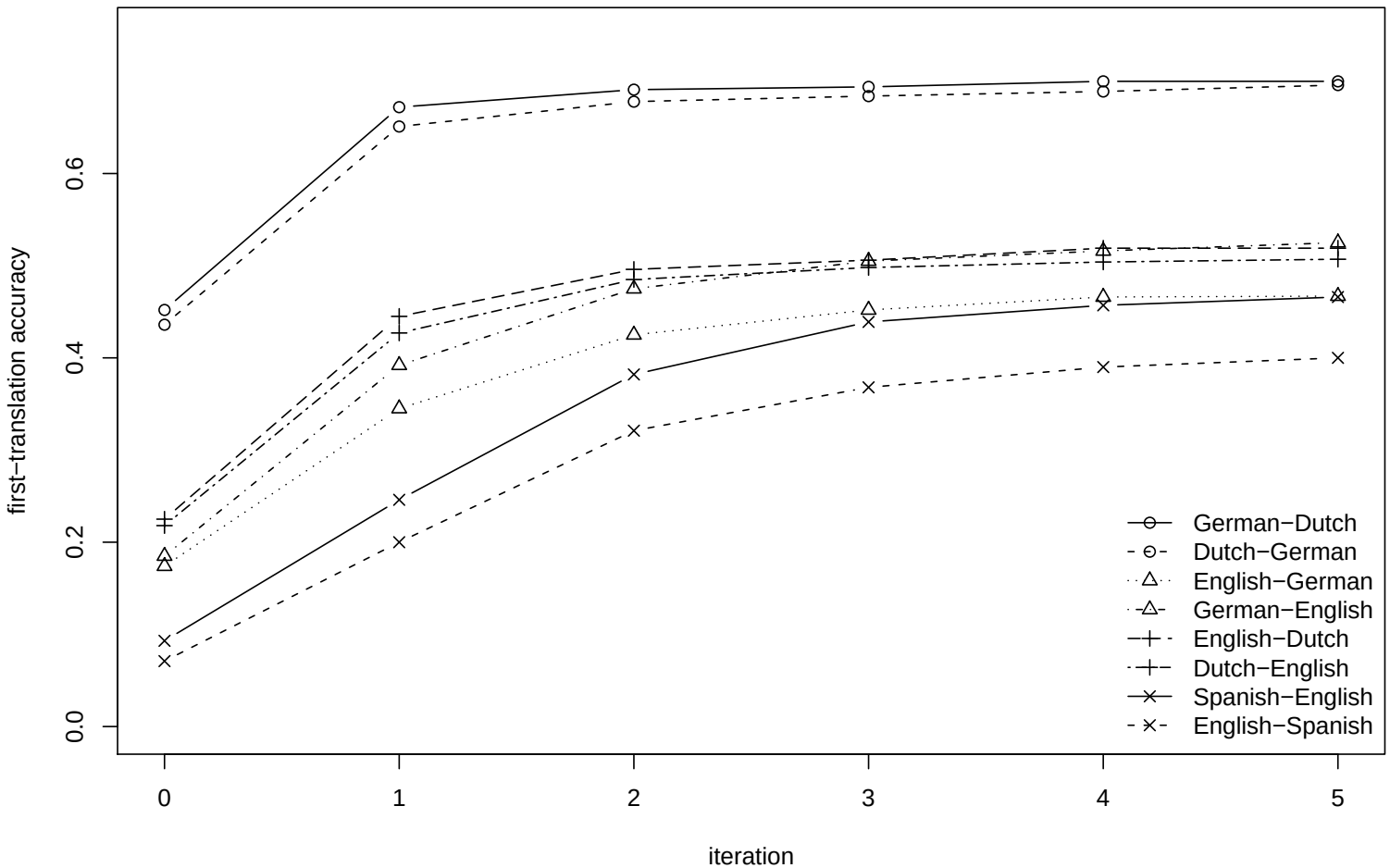
Experiments

- Experiments for Dutch-English, Dutch-German, German-English, English-Spanish
- Additional stemming for English-Spanish
- Final results

	n	noun	adj	verb	total
Dutch-German	3185	.710	.698	.659	.696
German-Dutch	3137	.732	.691	.633	.700
Dutch-English	3954	.559	.535	.359	.507
English-Dutch	3815	.582	.559	.339	.519
English-German	8042	.487	.538	.368	.467
German-English	7340	.562	.525	.427	.525
English-Spanish	6287	.423	.473	.268	.400
Spanish-English	5447	.491	.532	.327	.466

Cross-lingual similarity

Accuracy improves through the process



Cross-lingual similarity

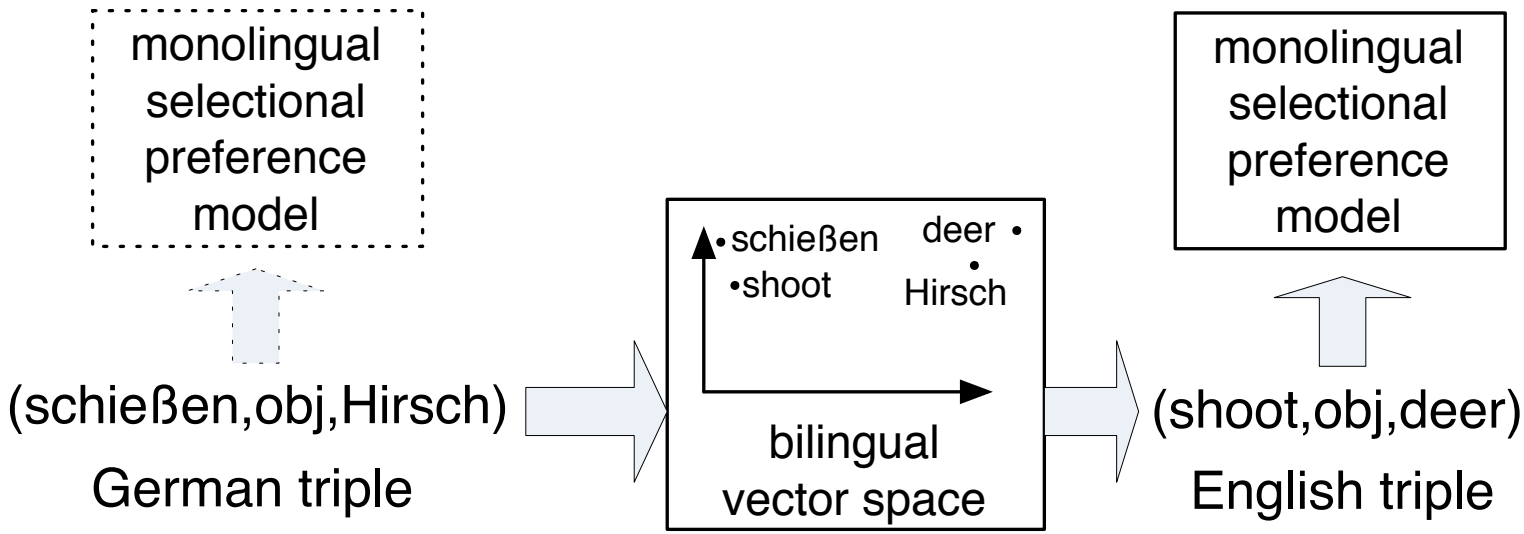
Examples

stad	=	city	∼	town	∼	area
roman	=	novel	∼	story	∼	poem
aap	∼	animal	∼	elephant	=	monkey
vrouw	=	woman	∼	man	∼	child
bedreigen	=	threaten	∼	protect	∼	attack
leren	=	learn	=	teach	∼	know
veronderstel	∼	imply	∼	regard	=	assume
wemel	?	translate	∼	abound	∼	crowd

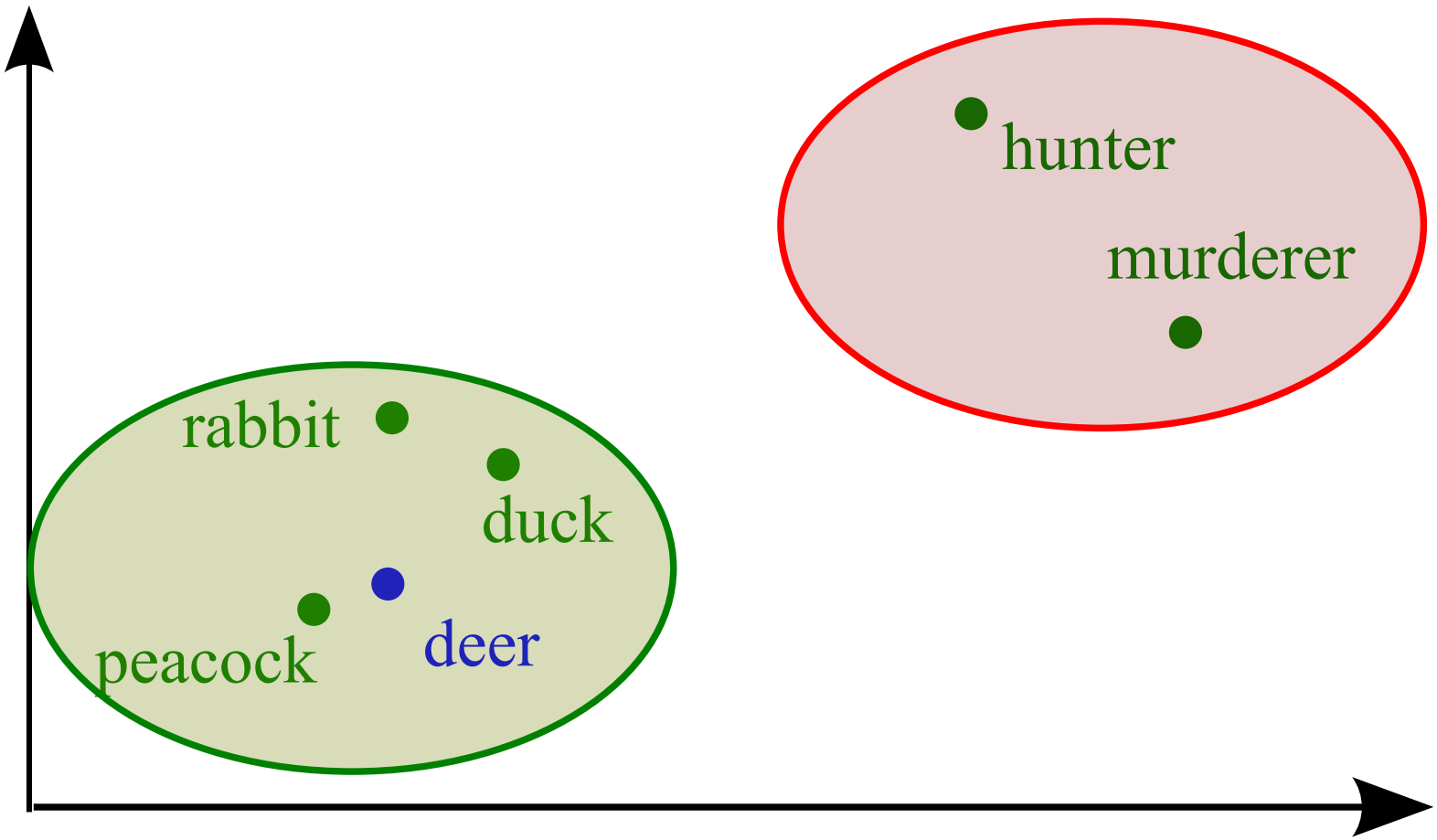
Cross-lingual similarity

Applications

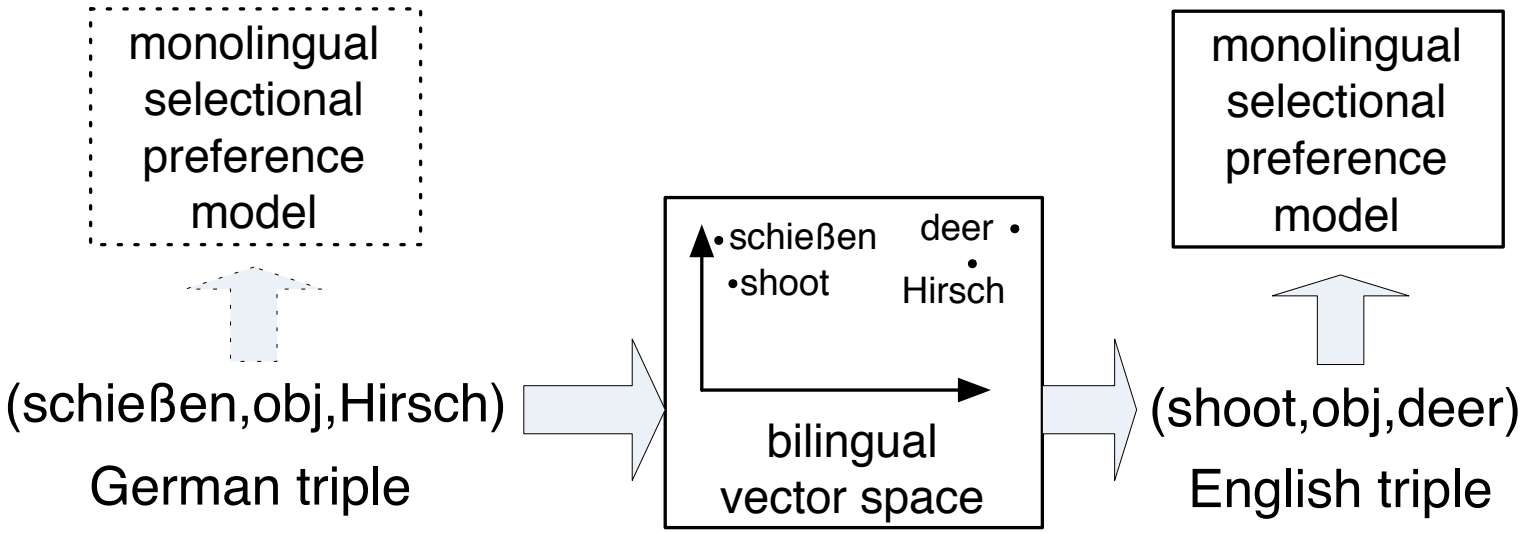
- Identification of false friends
- Bilingual knowledge induction



Cross-lingual similarity



Cross-lingual similarity



Outline

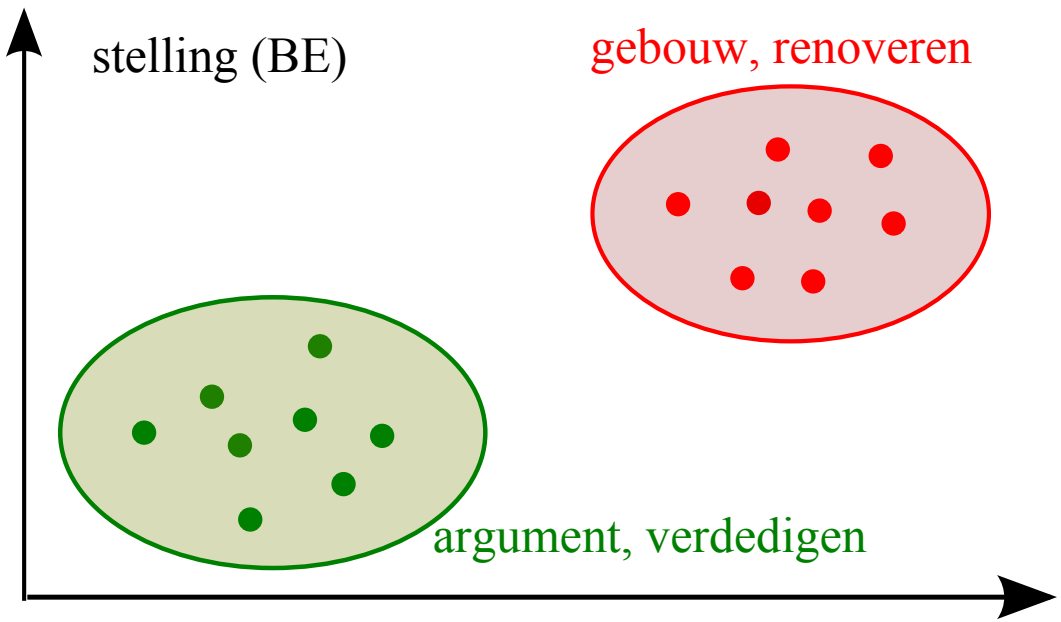
Conclusions and Outlook

- Extension of semantic similarity to two corpora
- Application to lexical variation
  - Identification of lectal markers
  - Identification of cross-lectal synonyms
- Application to bilingual knowledge induction
  - Basic translation of words between languages
  - Generalizing knowledge from one language to the other

Conclusions and Outlook

Polysemy

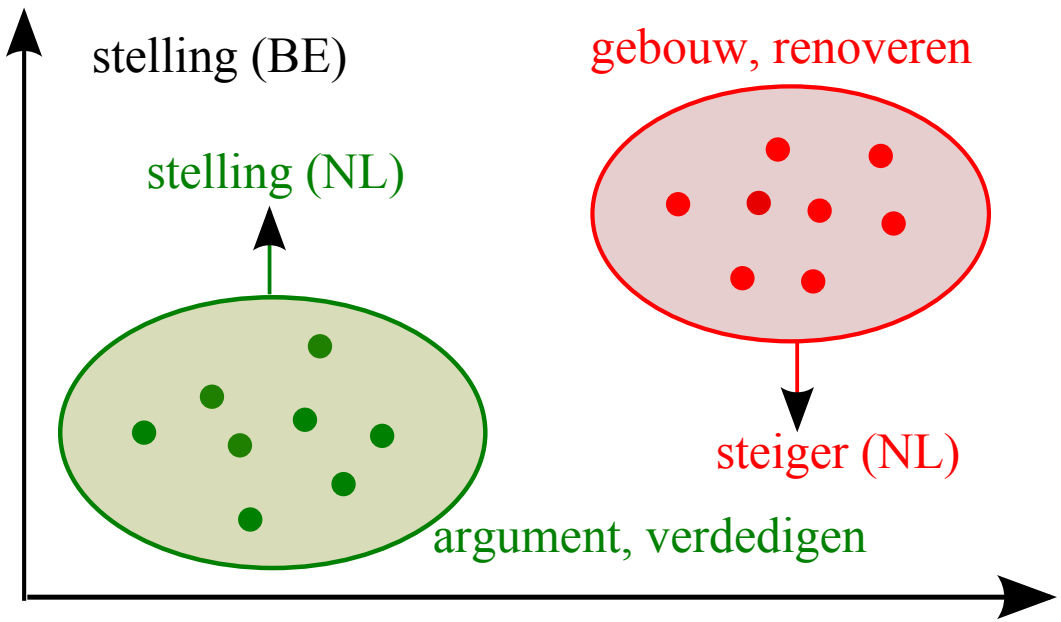
- Addressing polysemy involves a movement towards individual contexts
- Looking for a clustering of contexts that corresponds to the sense structure of a word



Conclusions and Outlook

Polysemy

- Addressing polysemy involves a movement towards individual contexts
- Looking for a clustering of contexts that corresponds to the sense structure of a word



For more information:

http://wwwling.arts.kuleuven.be/qlvl

yves.peirsman@arts.kuleuven.be