Modelling Meaning

Text-Based Analysis of Concept Structure and Lexical Semantics with Distributional Semantic Models

- Kris Heylen -

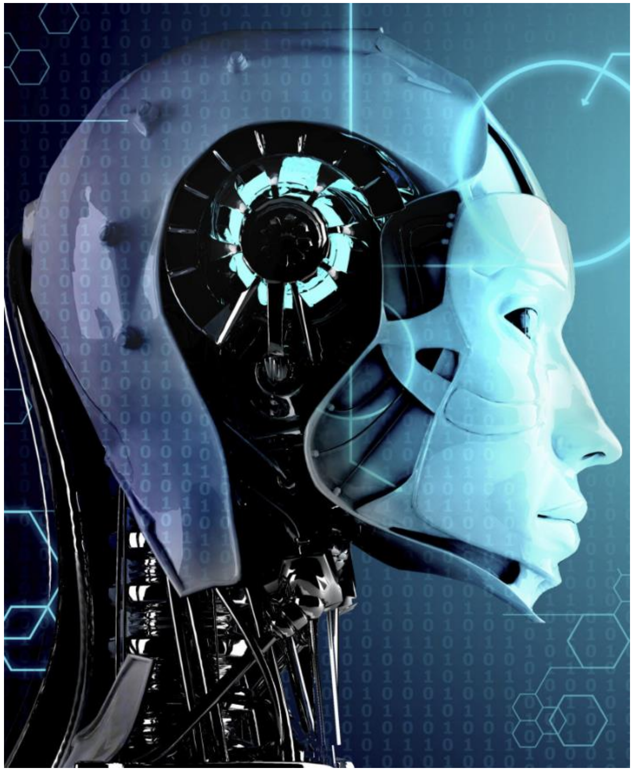
31 March 2017

KU LEUVEN

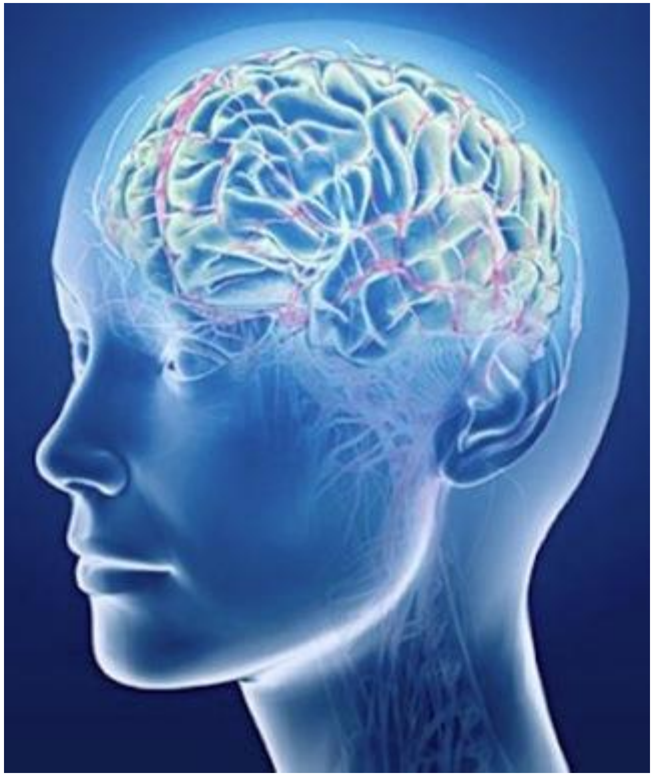
QLVL

VUB VRIJE UNIVERSITEIT BRUSSEL

Which AI approach?



Engineering

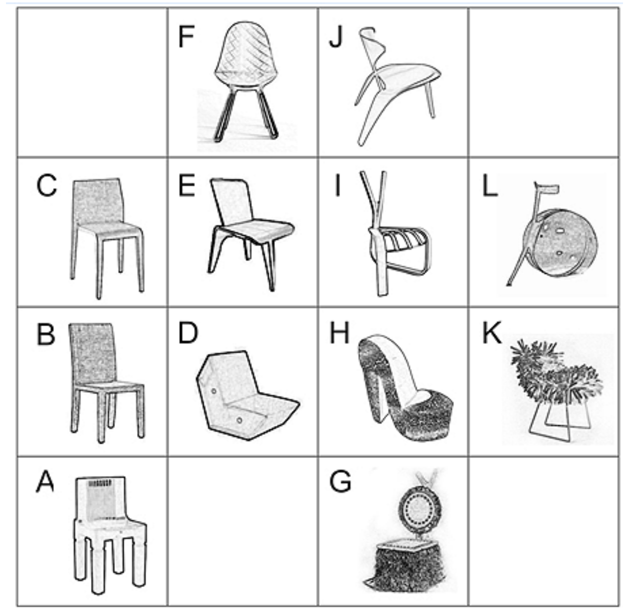


Cognitive science

Categorization

Cognitive psychology: How do we organize our experiences into clusters? (categories / concepts)

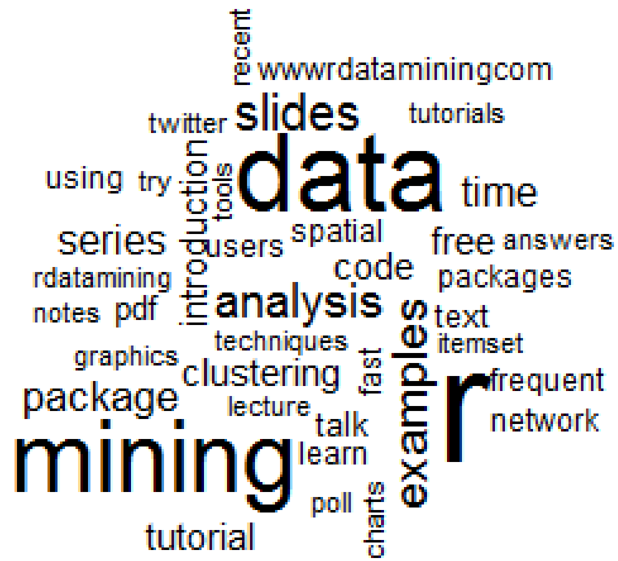
Classical approach: Necessary and sufficient conditions (logic)



Cognitively realistic approach:

- Prototype theory
- Exemplar theory

Referential vs discursive concepts



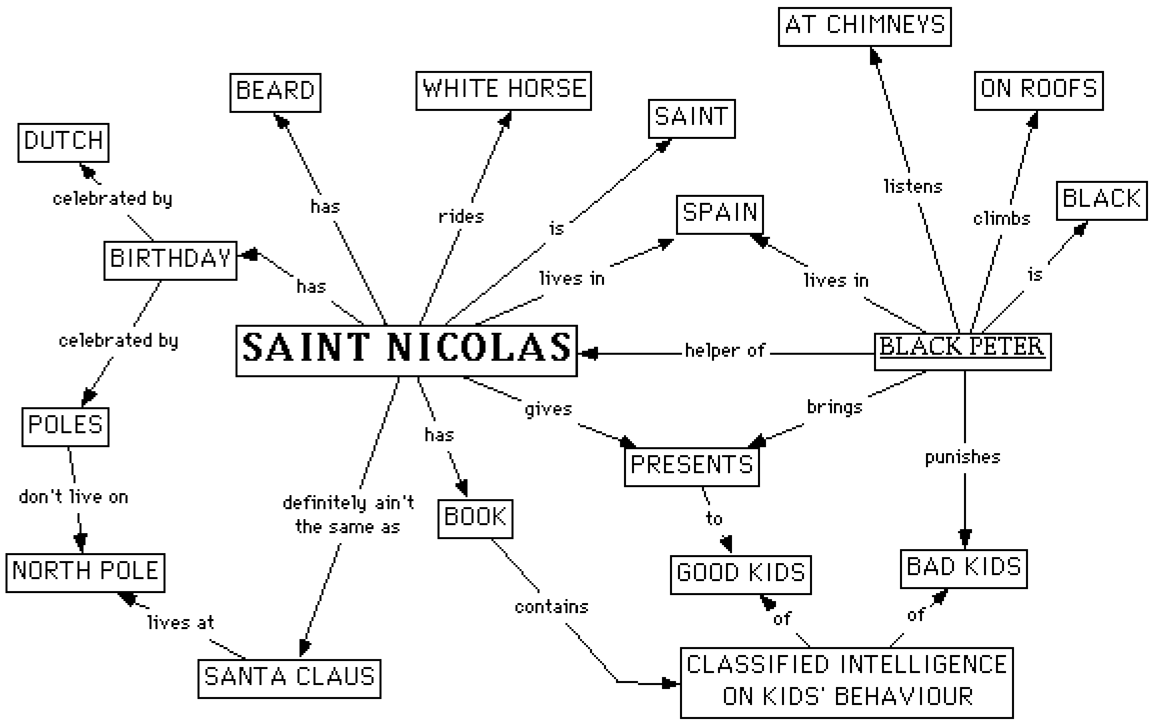
Sensory perception

Experience

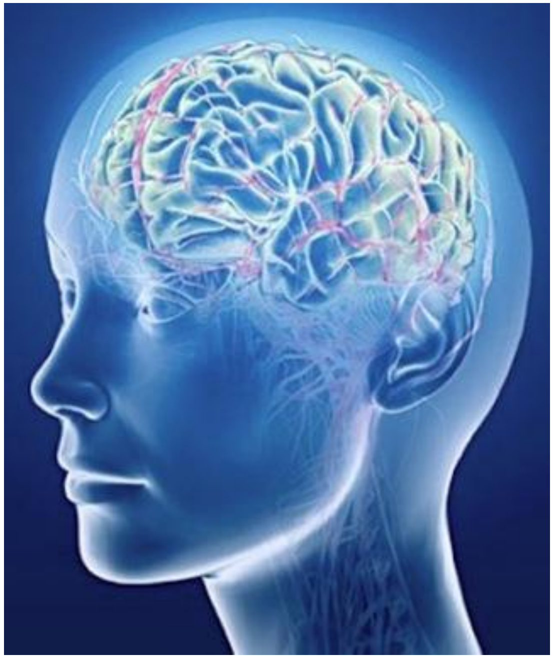
Reasoning & discourse

Language

Network of concepts



Cognitive and Social Phenomenon



Concepts are cognitively grounded in the brain

Concepts get shaped by human interaction

Overview

1. Concepts in Cognitive Sociolinguistics
2. Distributional Corpus Analysis
3. Case Study 1: Discourse about immigrants
4. Case Study 2: Positive evaluative adjectives
5. Conclusion

1. Cognitive Sociolinguistics

QLVL's research is situated in what has become known as Cognitive Sociolinguistics (Kristiansen & Dirven 2008; Geeraerts, Kristiansen& Peirsman 2010):

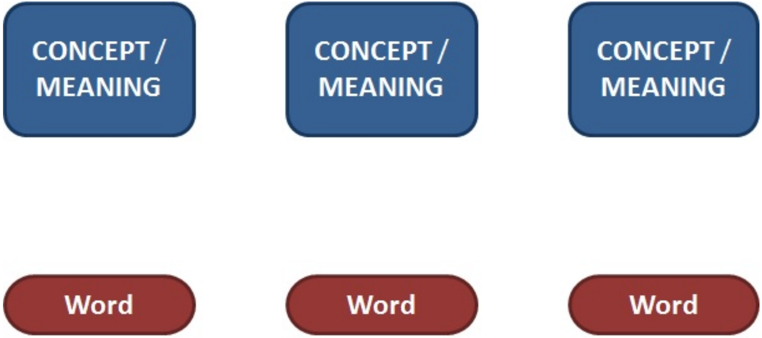
- a meaning-centered theory of language
- a usage-based perspective of language
- emphasis on the a socio-cultural aspects of semantic structure
- commitment to the use of advanced quantitative methods

QLVL has been developing this line of research since the 1990s:

- Structure of Lexical Variation (1994)
- Diachronic Prototype Semantics (1997)
- Profile-based approach (1999)

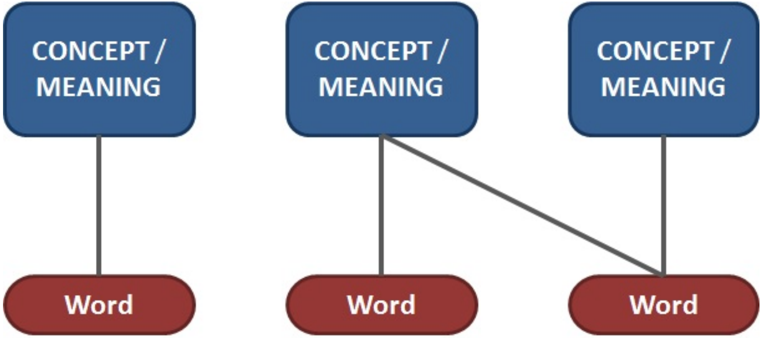
Structure of Lexical Variation (1994)

LEXICOLOGY (Geeraerts, Grondelaers & Bakema 1994):



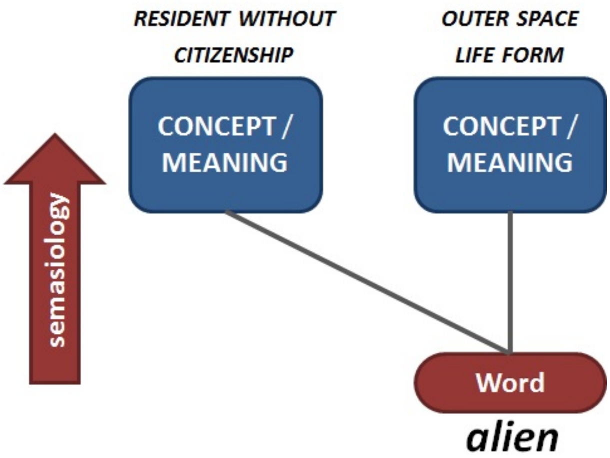
Structure of Lexical Variation (1994)

LEXICOLOGY (Geeraerts, Grondelaers & Bakema 1994):



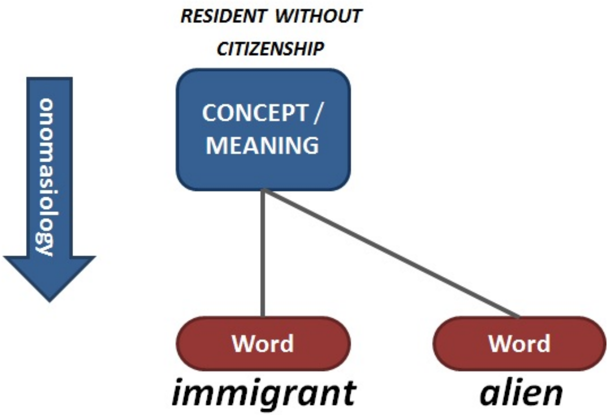
Structure of Lexical Variation (1994)

SEMASIOLOGY:



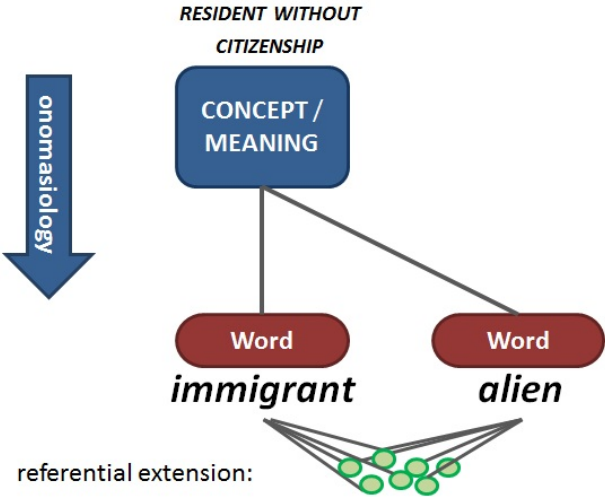
Structure of Lexical Variation (1994)

ONOMASIOLOGY:



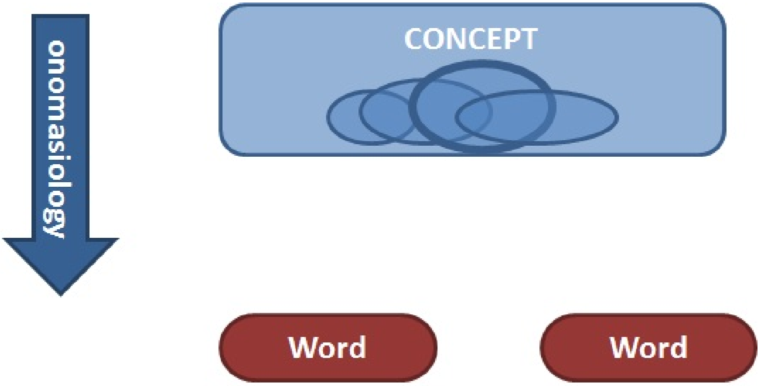
Structure of Lexical Variation (1994)

ONOMASIOLOGY:



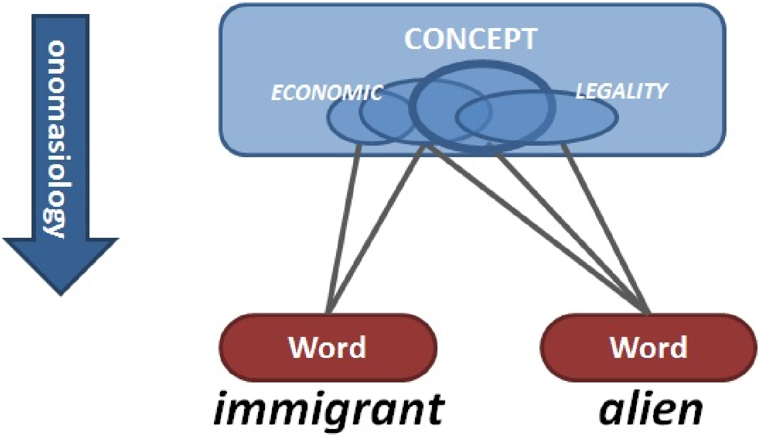
Structure of Lexical Variation (1994)

PROTOTYPE STRUCTURE:



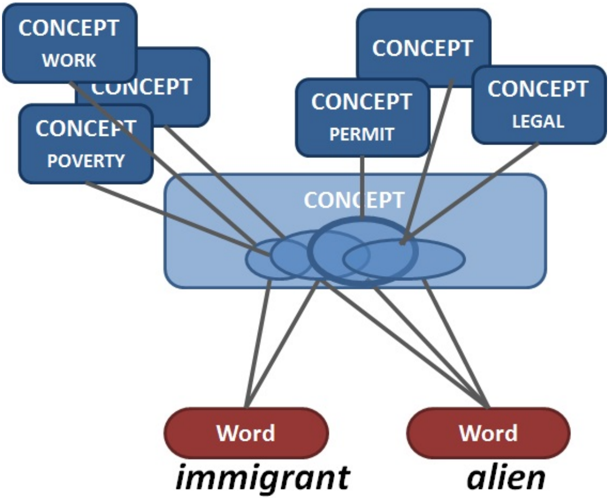
Structure of Lexical Variation (1994)

CONSTRUAL:



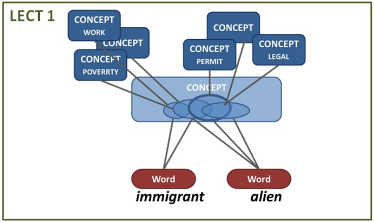
Structure of Lexical Variation (1994)

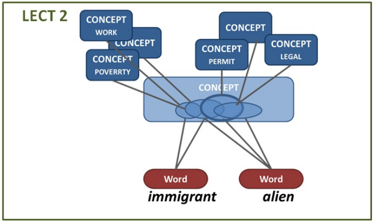
CONCEPT NETWORK:



Structure of Lexical Variation (1994)

LECTAL VARIATION:





GEOGRAPHY

REGISTER

MEDIUM

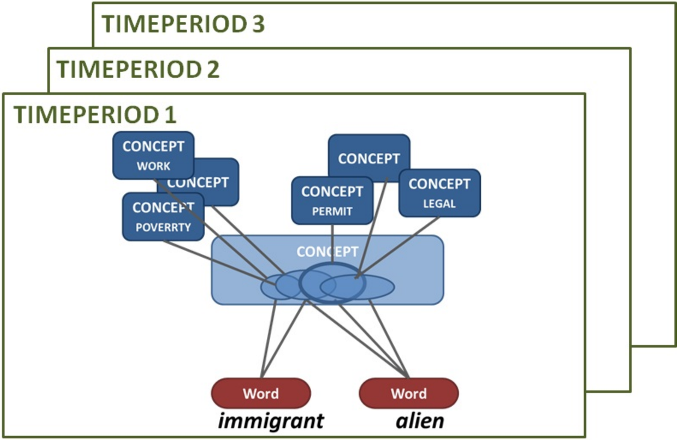
TIME

1966

2016

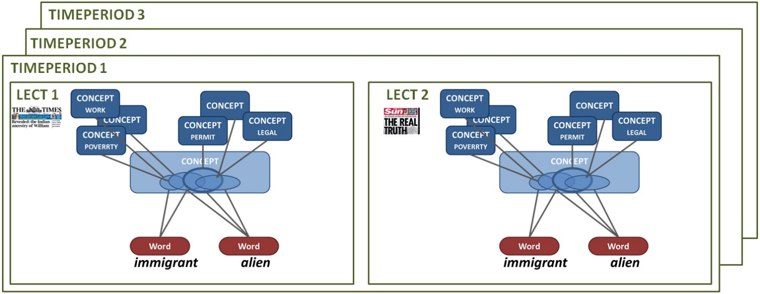
Structure of Lexical Variation (1994)

DIACHRONIC VARIATION:

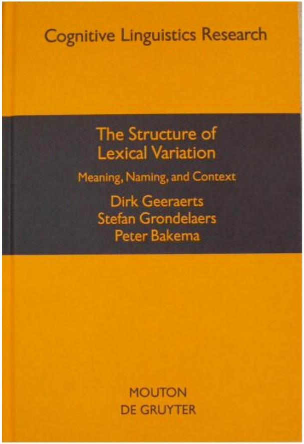


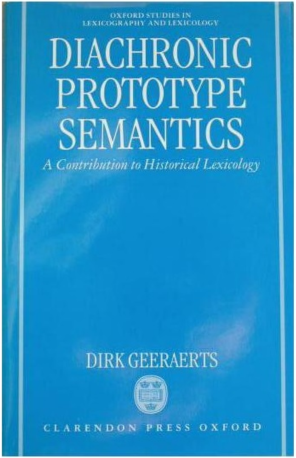
Structure of Lexical Variation (1994)

DIACHRONIC REGISTER VARIATION:



Structure of Lexical Variation (1994)



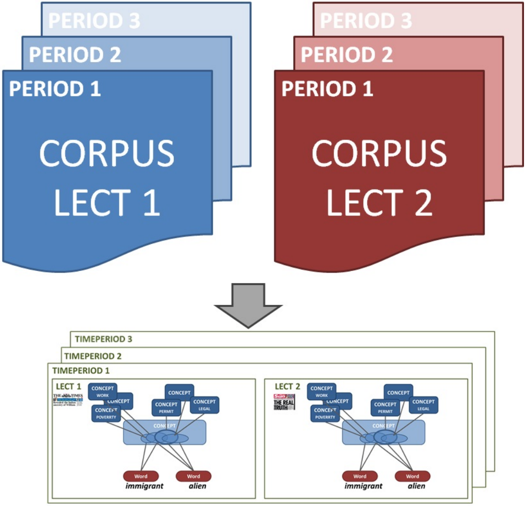


Overview

1. Concepts in Cognitive Sociolinguistics
2. Distributional Corpus Analysis
3. Case Study 1: Discourse about immigrants
4. Case Study 2: Positive evaluative adjectives
5. Conclusion

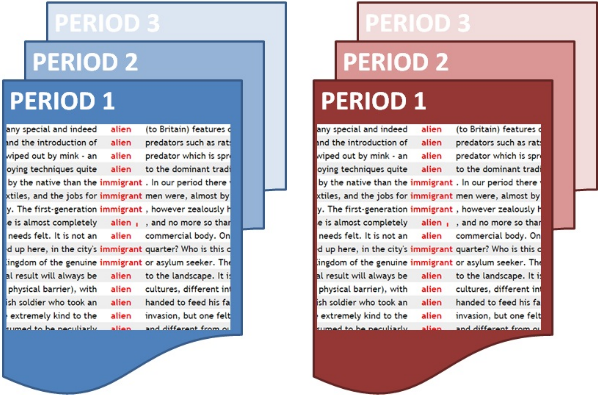
Corpus Analysis

STRATIFIED CORPORA



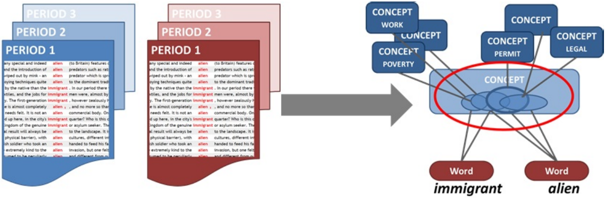
Corpus Analysis

CONCORDANCES



Corpus Analysis

METHODOLOGY FOR FINDING CONCEPTUAL FEATURES?



		FEATURE RETRIEVAL	
		MANUAL	STATISTICAL
FEATURE ANALYSIS	MANUAL	*Philology*	*Sinclair-style corpus Lx*
	STATISTICAL	*Behavioral Profiles*	***Distributional models***

Distributional semantic modelling

Linguistic origin: Distributional Hypothesis

- "You shall know a word by the company it keeps" (Firth)
- a word's meaning can be induced from its co-occurring words

Semantic Vector Spaces in Computational Linguistics

- standard technique in statistical NLP for the large-scale automatic modeling of (lexical) semantics
- aka Vector Spaces Models, Distributional Semantic Models, Word Spaces,... (cf Turney & Pantel 2010 for overview)
- generalised, large-scale collocation analysis
- words occurring in same contexts have similar meaning

Semantic Vector Spaces as models of word meaning

Practical

Which two words out of a set of three have the same meaning?

ongeval, koffie, accident

Occurrences in context from a corpus

Op de Brusselse ring deed zich een	ongeval	met een vrachtwagen voor
's Morgens drinkt hij een kop	koffie	met melk en suiker
2 bestuurders raakten gekwetst bij een	ongeval	met een vrachtwagen
in de avondspits veroorzaakte een	accident	een kilometerslange file
als vieruurtje serveert het hotel	koffie	en gebak voor de gasten
de auto was betrokken in een	accident	met een dodelijke afloop
Met winterbanden is het risico op een	ongeval	bij vriesweer veel kleiner

	auto	slachtoffer	vrachtwagen	file	gekwetst	suiker	melk	kop
ongeval	0	0	0	0	0	0	0	0

	auto	slachtoffer	vrachtwagen	file	gekwetst	suiker	melk	kop
ongeval	0	0	1	0	1	0	0	0

vader raakte gekwetst bij een ongeval met een vrachtwagen op de

	auto	slachtoffer	vrachtwagen	file	gekwetst	suiker	melk	kop
ongeval	0	0	1	1	1	0	0	0

voor zeven uur veroorzaakte een ongeval een kilometerslange file richting Antwerpen

	auto	slachtoffer	vrachtwagen	file	gekwetst	suiker	melk	kop
ongeval	0	1	2	1	1	0	0	0

vrachtwagens waren betrokken bij het ongeval, dat meer dan tien slachtoffers

	auto	slachtoffer	vrachtwagen	file	gekwetst	suiker	melk	kop
ongeval	1	2	2	1	1	0	0	0

	auto	slachtoffer	vrachtwagen	file	gekwetst	suiker	melk	kop
ongeval	1	3	2	1	2	0	0	0

	auto	slachtoffer	vrachtwagen	file	gekwetst	suiker	melk	kop
ongeval	1	4	2	1	4	0	0	0

	auto	slachtoffer	vrachtwagen	file	gekwetst	suiker	melk	kop
ongeval	37	83	142	17	66	0	0	0

	auto	slachtoffer	vrachtwagen	file	gekwetst	suiker	melk	kop
ongeval	120	424	388	82	270	11	3	1

	auto	slachtoffer	vrachtwagen	file	gekwetst	suiker	melk	kop
ongeval	120	424	388	82	270	11	3	1
accident	0	1	2	2	0	0	0	0

	auto	*slachtoffer*	*vrachtwagen*	*file*	*gekwetst*	*suiker*	*melk*	*kop*
ongeval	120	424	388	82	270	11	3	1
accident	1	3	2	4	1	0	0	0

	auto	slachtoffer	vrachtwagen	file	gekwetst	suiker	melk	kop
ongeval	120	424	388	82	270	11	3	1
accident	10	17	22	7	0	0	0	1

	auto	slachtoffer	vrachtwagen	file	gekwetst	suiker	melk	kop
ongeval	120	424	388	82	270	11	3	1
accident	53	121	67	24	55	2	0	3

	auto	slachtoffer	vrachtwagen	file	gekwetst	suiker	melk	kop
ongeval	120	424	388	82	270	11	3	1
accident	154	401	376	99	305	20	1	5

	auto	slachtoffer	vrachtwagen	file	gekwetst	suiker	melk	kop
ongeval	120	424	388	82	270	11	3	1
accident	154	401	376	99	305	20	1	5
koffie	0	0	0	0	0	1	2	2

	auto	slachtoffer	vrachtwagen	file	gekwetst	suiker	melk	kop
ongeval	120	424	388	82	270	11	3	1
accident	154	401	376	99	305	20	1	5
koffie	0	0	0	0	0	3	5	4

	auto	slachtoffer	vrachtwagen	file	gekwetst	suiker	melk	kop
ongeval	120	424	388	82	270	11	3	1
accident	154	401	376	99	305	20	1	5
koffie	0	0	2	0	0	16	24	21

	auto	slachtoffer	vrachtwagen	file	gekwetst	suiker	melk	kop
ongeval	120	424	388	82	270	11	3	1
accident	154	401	376	99	305	20	1	5
koffie	3	5	11	1	0	55	76	64

	auto	slachtoffer	vrachtwagen	file	gekwetst	suiker	melk	kop
ongeval	120	424	388	82	270	11	3	1
accident	154	401	376	99	305	20	1	5
koffie	5	8	18	4	1	72	102	93

Which words are similar?

Distributional models of lexical semantics

word by word similarity matrix

	ongeval	accident	koffie
ongeval	1	.91	.08
accident	.91	1	.17
koffie	.08	.17	1

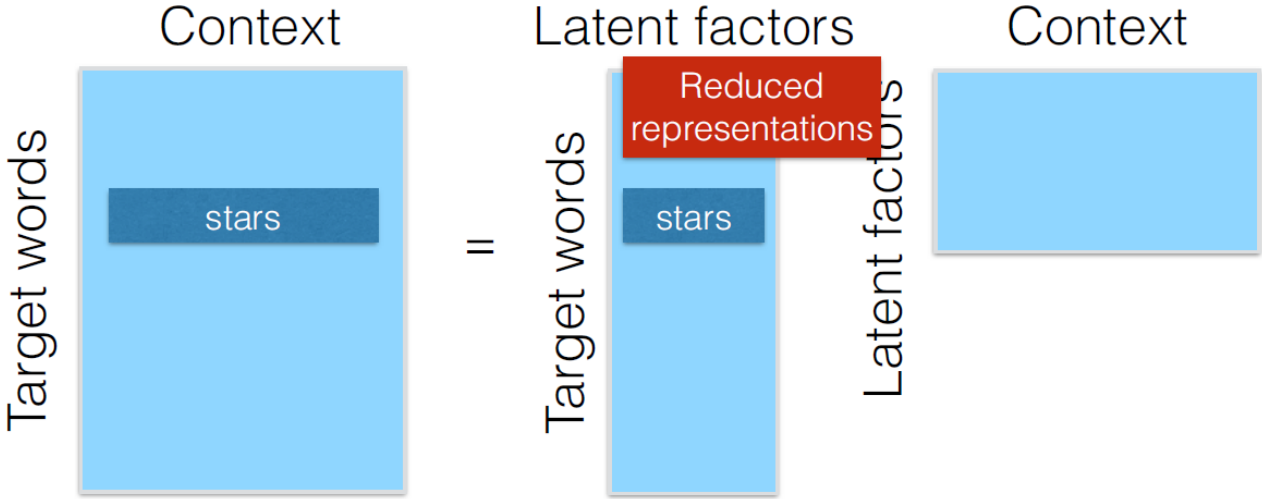
Point-wise mutual information (PMI)

$$\mathrm{PMI}(target, ctxt) = \log \frac{\mathrm{P}(target, ctxt)}{\mathrm{P}(target)\mathrm{P}(ctxt)}$$

	…	bright	in	sky	…	
stars	…	80	300	61	…	Raw counts
stars	…	3.1	1.2	2.4	…	PMI scores

- Other weighting schemes:
  - Tf-idf, Local mutual information, Log-Likelihood Ratio

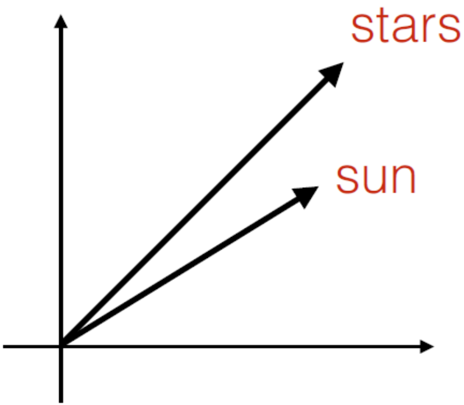
Dimensionality reduction



Factorize the co-occurrence counts as linear combinations over latent factors

From vectors to similarity in meaning

1. Extract co-occurrence counts
2. Apply a re-weighting scheme on the resulting co-occurrence matrix
3. Apply dimensionality reduction
4. Vector similarity



Cosine similarity

$$cos(\vec{u}, \vec{v}) = \frac{\Sigma_i u_i v_i}{\sqrt{\Sigma_i u_i^2}\sqrt{\Sigma_i v_i^2}}$$

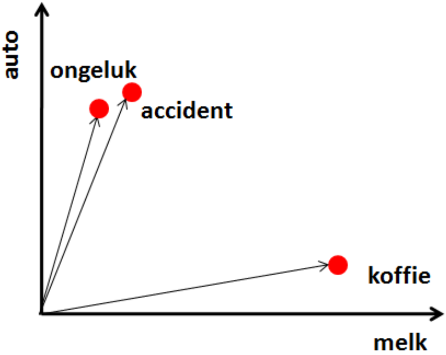
$$= \frac{< u, v >}{||u|| \times ||v||}$$

Other similarity measures: Euclidean, Lin

Distributional models of lexical semantics

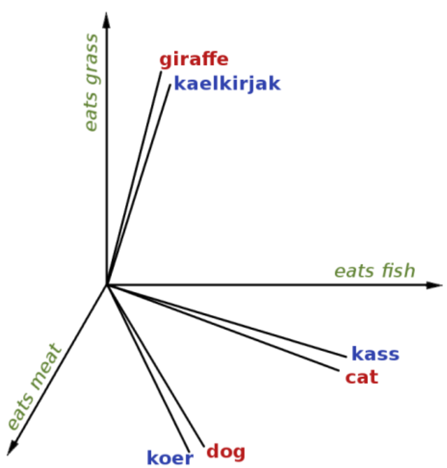
Geometrical metaphor: Semantic distance

- frequencies weighted by collocational strength (pmi)
- vectors projected in context feature space: Word Space
- cosine of angle between vectors as semantic similarity measure

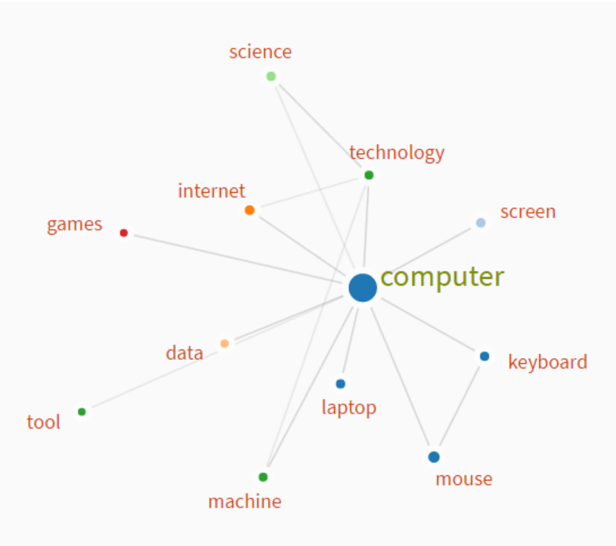


Computational Representations

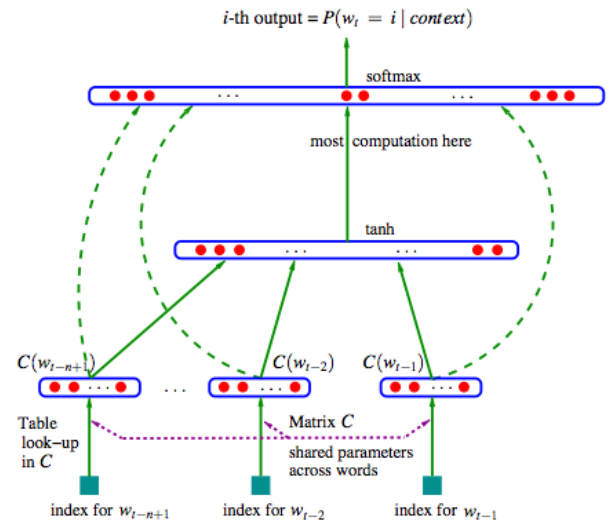
Vectors



Graphs



Neural Nets



Semantic neighbours of words

rhino	fall	good
woodpecker	rise	bad
rhinoceros	increase	excellent
swan	fluctuation	superb
whale	drop	poor
ivory	decrease	improved
plover	reduction	perfect
elephant	logarithm	clever
bear	decline	terrific

http://clic.cimec.unitn.it/infomap-query/

Semantic neighbours of phrases

DIRT - Lin and Pantel, 2007

X's addiction to Y

Cosmos

N:gen:N<addiction>N:to:N

1 N:gen:N<**addiction**>N:nn:N
2 N:gen:N<**craving**>N:for:N
3 N:gen:N<**child**>N:about:N
4 N:gen:N<**money**<N:obj:V<**spend**>V:on:N
5 N:gen:N<**intake**>N:nn:N
6 N:gen:N<**zest**>N:for:N
7 N:gen:N<**winning**>N:nn:N
8 N:gen:N<**use**>N:nn:N
9 N:gen:N<**habit**>N:nn:N

X manufactures Y

Cosmos

N:subj:V<manufacture>V:obj:N

1 N:by:V<**manufacture**>V:obj:N
2 N:obj:V<**manufacture**>V:subj:N
3 N:subj:V<**produce**>V:obj:N
4 N:subj:V<**begin**>V:obj:N>**production**>N:of:N
5 N:subj:V<**export**>V:obj:N
6 N:subj:N<**supplier**>N:of:N
7 N:subj:V<**supply**>V:obj:N
8 N:subj:V<**sell**>V:obj:N
9 N:appo:N<**manufacturer**>N:nn:N

http://demo.patrickpantel.com/demos/lexsem/paraphrase.htm

General-purpose representations of meaning

- Synonymy
- Relatedness
- Concept categorization
- Selectional preferences
- Analogy
- Relation classification
- …

Similarity/relatedness

- WordSim-353, SimLex-999, MEN

chapel	church	0.45
eat	strawberry	0.33
jump	salad	0.06
bikini	pizza	0.01

- Evaluation: Correlation of model cosines with human similarity assessments (close to human performance on relatedness, difficulties on synonym detection)

Selectional preferences

- Pado 2007

eat	villager	obj	1.7
eat	pizza	obj	6.8
eat	pizza	subj	1.1

- Evaluation: Create prototype argument vector (average all OBJ vectors of *eat*), compute similarity of prototype with candidate argument (*pizza*)

Categorization

- ESSLLI 2008 Shared task, Almuhareb and Poesio 2006

VEHICLE	MAMMAL
helicopter	dog
motorcycle	elephant
car	cat

- Evaluation: Cluster word vectors, overlap between clusters and gold categories (close to 90% cluster purity with 6 categories)

Distributional semantics: some references

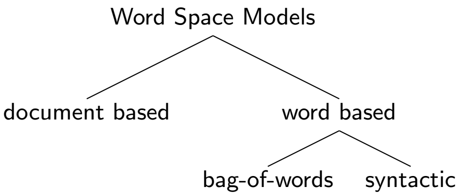
- Overviews
  - Turney and Pantel 2010, Pado and Lapata 2007, Erk 2012, Baroni, Bernardi, Zamparelli - Frege in Space 2014
- Comparisons/evaluation
  - Agirre et al, 2009, Baroni and Lenci 2010, Bullinaria and Levy 2007, Bullinaria and Levi2012, Sahlgren 2006, Kiela et al 2014

Different Context Models

Vector Space Models come in many flavours

The main difference between the models lies in how they define context

Family of models



Different Context Models

document based models

- context = stretch of text in which target word occurs
- 2 words are related when they often co-occur in text
- Landauer & Dumais 1997: Latent Semantic Analysis

word based models

- context = context words around the target word
- 2 words are related when they co-occur with the same context words, but not necessarily with each other

	DOC.1	DOC.2	DOC.3	DOC.4	DOC.5	DOC.6	DOC.7	DOC.8
ongeval	23	12	14	24	8	0	0	0
ongeluk	16	9	11	18	17	20	0	1
koffie	0	0	0	0	0	14	12	15

	auto	slachtoffer	vrachtwagen	file	gekwetst	suiker	melk	kop
ongeval	120	424	388	82	270	11	3	1
ongeluk	154	401	376	99	305	20	1	5
koffie	5	8	18	4	1	72	102	93

Different Context Models

Within word based models:

bag-of-words

- context words in window of *n* words left and right of target word
- a bag of unstructured context features

syntactic features

- context words in specific syntactic relation with target word
- takes clause structure into account
- Lin 1998, Padó & Lapata 2007

The wagging dog barked at the postman on the bike

	wagging	dog	bark	postman	bike
dog	1	0	1	1	1
postman	1	1	1	0	1

The wagging dog barked at the postman on the bike

	=subj.bark	+adj.wagging	=PC.bark.at	+PP.on.bike
dog	1	1	0	0
postman	0	0	1	1

Different Context Models

Within the bag-of-words models:

1st order co-occurrences

- context = words in immediate proximity to the target
- Levy & Bullinaria 2001

2nd order co-occurrences

- context = context words of context words of target
- can generalise over semantically related context words
- Schütze 1998

NB syntactic models are also 1st order models

Introduction

Distribution over subcategorization frames

- close link to Structuralism (Apresjan, Levin)
- context feature = combination of verb arguments
- in principle purely syntactic information
- task: verb classifcation (e.g. Schulte im Walde 06, Joanis 07)
- Subcat Frame Models

Distribution over co-occurrences

- Distributional hypothesis (Harris, Firth, Sinclair)
- context feature = one co-occurring lexeme
- task: thesaurus extraction (e.g. Lin 98)
- Word Space Models

Subcat Frame Models: Subcategorization frames

	SU	SU/DO	SU/PC	SU/DO/PC	SU/DO/IO	DO/IO	SU/IO	SU/IO/CCI
fly	231	121	141	198	8	0	0	0
tell	1	221	0	88	301	12	4	25

Word Space Models: Lexical co-occurrences

	pilot	plane	bird	bee	airport	dog	cup	milk
fly	120	424	388	82	270	11	3	1
drink	1	21	25	2	16	19	323	401

Introduction

Continuum between syntactic and lexical features

- subcategorization approaches can take into specific prepositions or high level semantic information (Schulte im Walde 06)
- co-occurrence approaches take into account (complex) dependency relations (Pado & Lapata 07)
- approaches that combine subcategorization and lexical co-occurrence (Li and Brew 08, Van De Cruys 08)

lexically enriched subcategorization frames

	SUanim	SU/DO	SU/LC-over	SU/DO/LC-to	SU/DO/IO	DO/IO	SU/IO	SU/CC
fly	231	121	141	198	8	0	0	0

dependency based WSM: syntactically conditioned lexical co-occurrences

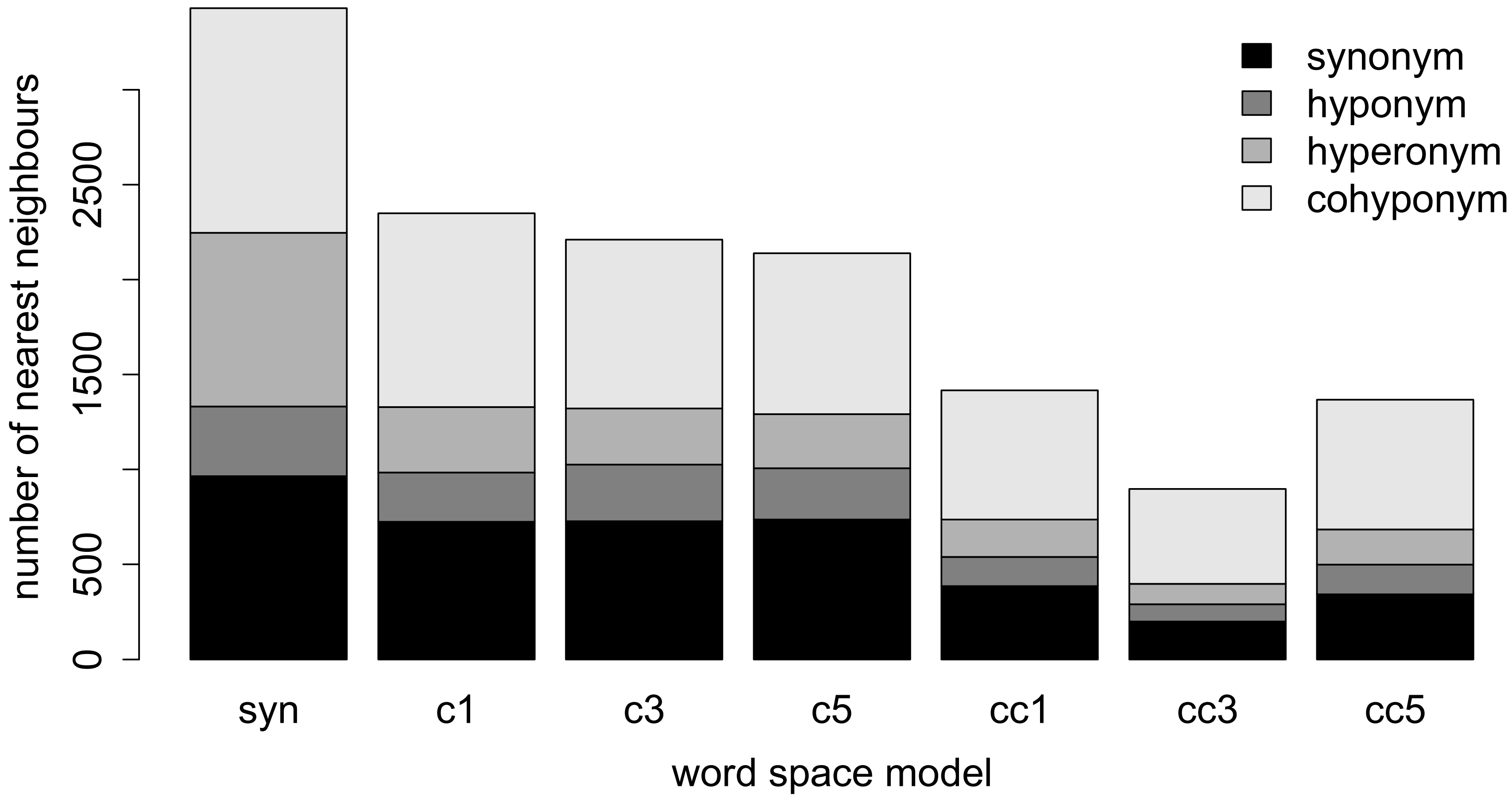
	SU pilot	DO plane	SU bird	SU bee	PC airport	SU dog	DO cup	PP milk
fly	120	424	388	82	270	11	3	1

Application 1: Finding near synonyms

English	Items
Manner	wijze, manier
Genocide	volk_moord, genocide
Poll	peiling, opiniepeiling
Marihuana	cannabis, marihuana
Putsch	staatsgreep, coup
Meningitis	hersenvliesontsteking, meningitis
Demonstrator	demonstrant, betoger
Airport	vliegveld, luchthaven
Coldness	koude, kou
Torture	marteling, foltering
Victory	zege, overwinning
Homosexual	homo, homoseksueel
Saxophone	sax, saxofoon
Internetprovider	provider, internetprovider, internetaanbieder
Airconditioning	airconditioning, airco
Religion	religie, godsdienst
The other side	overkant, overzijde
Explosion	explosie, ontploffing

Application 1: Finding near synonyms

Different context models (cf. infra)



Application 2: Lectometry

Profile-based approach

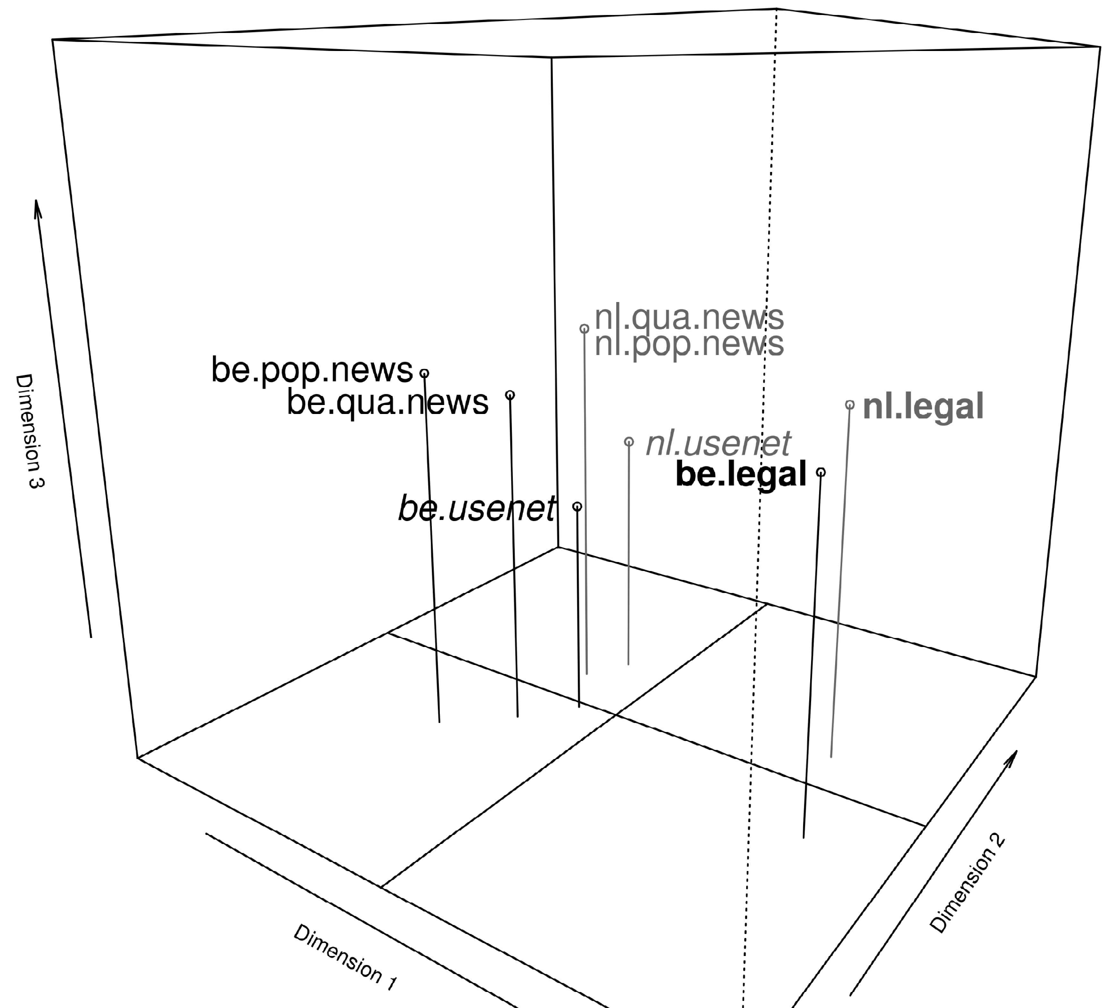
One concept JEANS:

	B	NL	overlap
jeans	85	30	30
spijkerbroek	15	70	15
			45

⇒ Aggregate over many distributionally generated profiles/synsets

Application 2: Lectometry

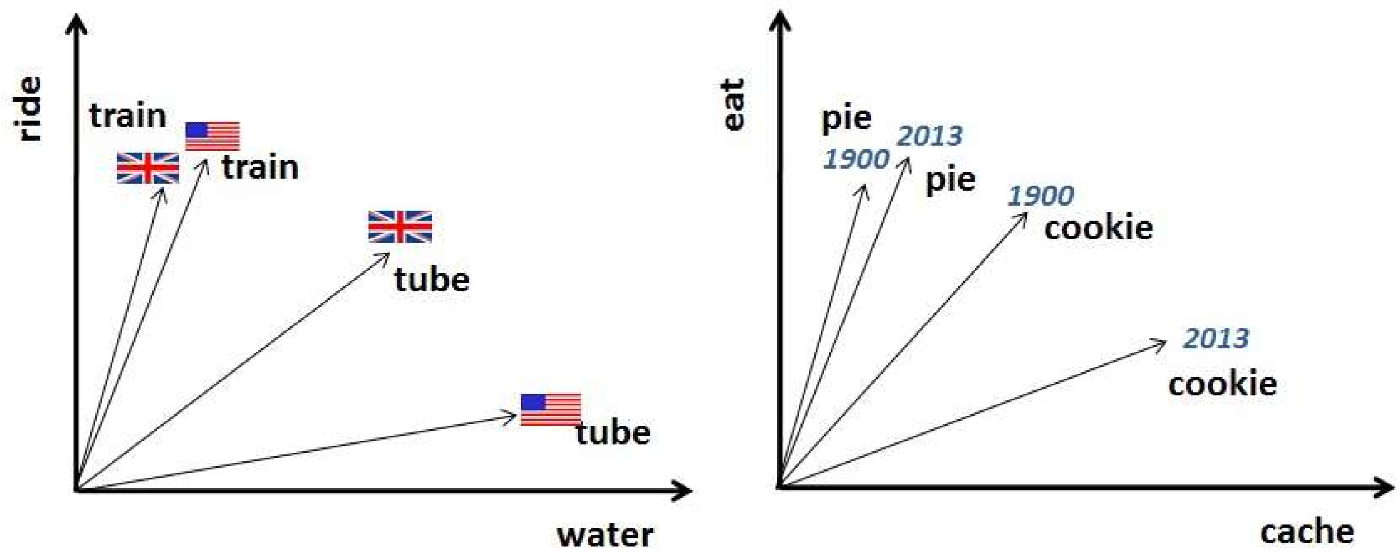
Profile-base approach (Ruette)



Application 3: Lexical variation

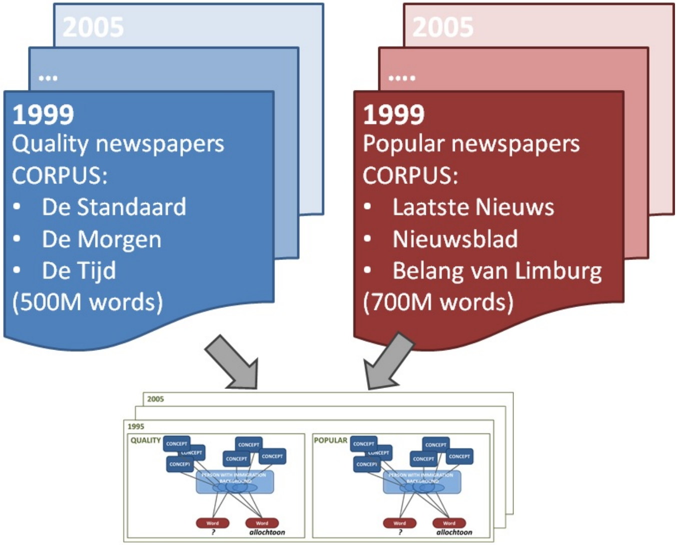
Bilectal Word Spaces (Peirsman)

- Extend Word Space from one corpus to two corpora representative for different lects/varieties
- 2 context vectors for each word, one for each variety
- most words will have themselves as most similar word, BUT words with diverging semantic structure will not



Overview

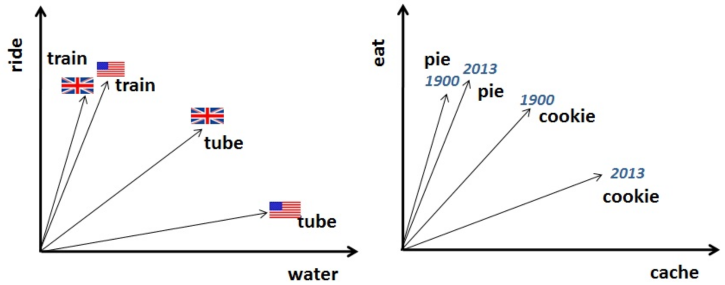
Concept IMMIGRANT in Belgian Newspapers



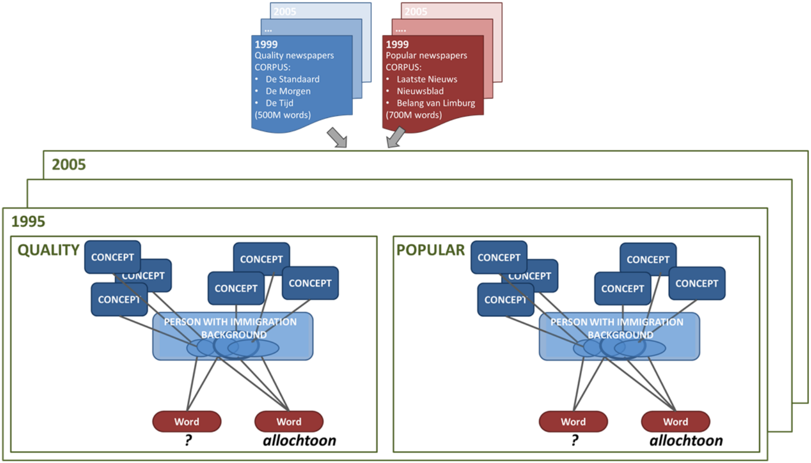
Distributional semantics: lexical variation

Bilectal Word Spaces

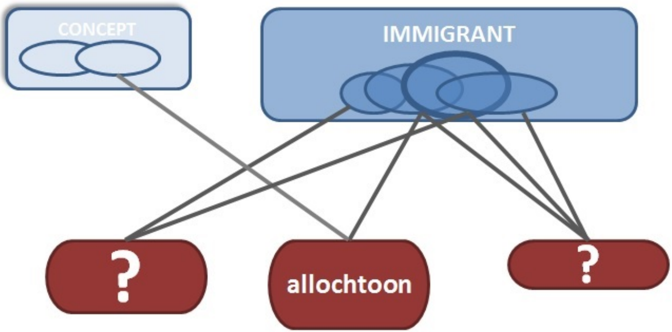
- Extend Word Space from one corpus to two corpora representative for different lects/varieties
- 2 context vectors for each word, one for each variety
- most words will have themselves as most similar word...
- BUT words with diverging semantic structure will not



Concept IMMIGRANT in Belgian Newspapers



Identifying alternative expressions



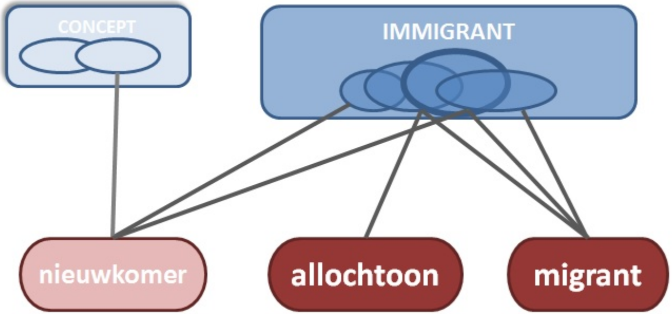
- calculate contextual similarity between 10K Dutch nouns
- sort by similarity to *allochtoon*

Identifying alternative expressions

allochtoon	1.0
migrant	0.71
vreemdeling	0.48
immigrant	0.47
buitenlander	0.47
nieuwkomer	0.32
gastarbeider	0.29

Table alternatives to *allochtoon*

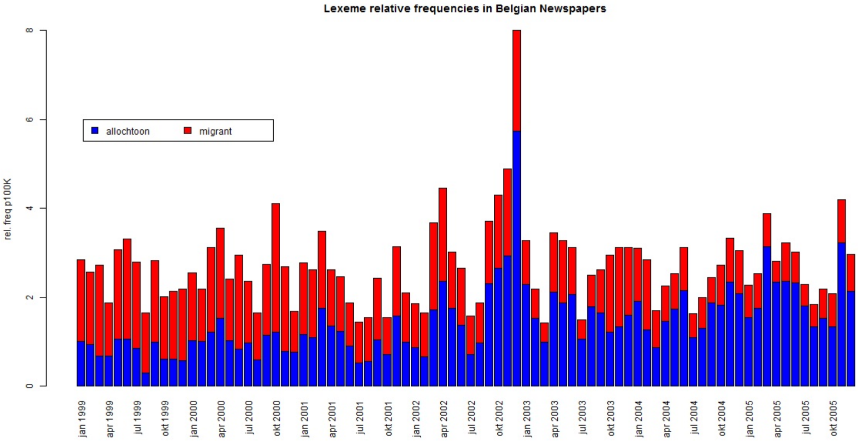
Identifying alternative expressions



Identifying alternative expressions

Normalised frequency of *allochtoon* and *migrant* per month

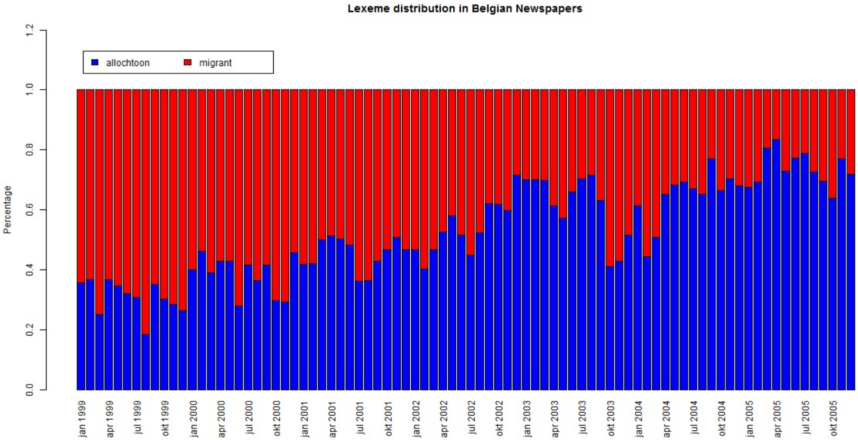
immigrant-talk seems to be a seasonal phenomenon



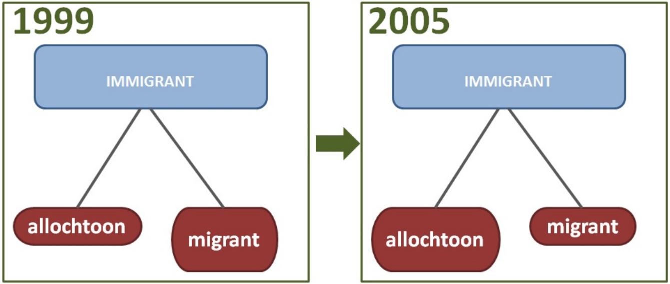
Identifying alternative expressions

Proportion of *allochtoon* and *migrant* in the corpus per month

allochtoon becomes more frequent than *migrant*



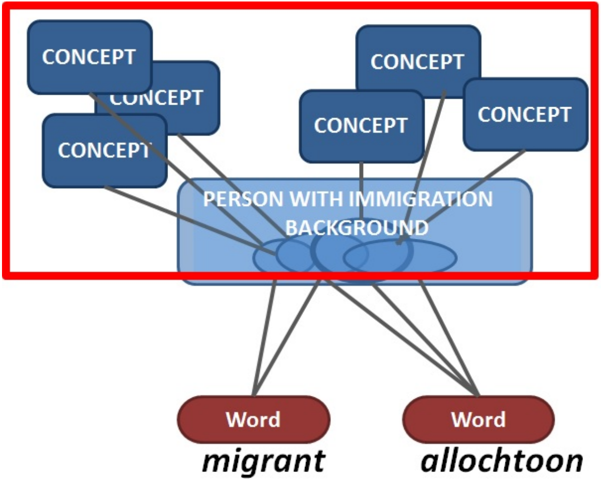
Identifying alternative expressions



Is this change in frequency also indicative of semantic change?

Analysing Semantic Structure

Which semantic features make up the internal structure of the concept?



Analysing Semantic Structure

Extract strongest concept collocations from matrix

	jobs	racisme	integratie	misdaad	stemrecht	suiker	zon	hond
allochtoon	5.3	7.9	6.5	4.0	0.8	0.6	0.0	0.0
migrant	4.3	8.1	5.7	3.2	6.2	0.5	0.0	0.1

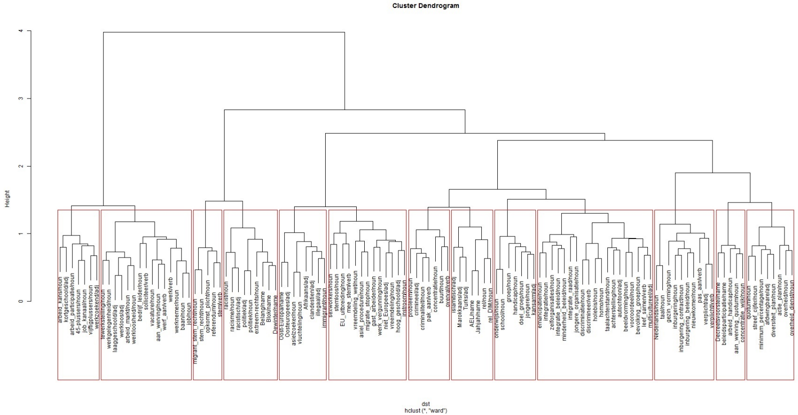
Analysing Semantic Structure

Make weighted co-occurrence matrix for these collocations

	jobs	racisme	integratie	misdaad	stemrecht	suiker	zon	hond
jobs	5.3	7.9	6.5	4.0	0.8	0.6	0.0	0.0
racisme	4.3	8.1	5.7	3.2	6.2	0.5	0.0	0.1
integratie	5.3	7.9	6.5	6.0	0.8	0.6	0.1	0.0
misdaad	4.3	8.1	5.7	2.2	6.2	0.4	0.0	0.1
stemrecht	5.3	7.9	6.5	8.0	0.8	0.9	0.3	0.0

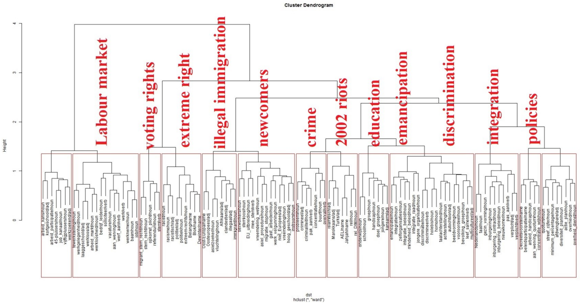
Analysing Semantic Structure

Calculate similarity between collocations and feed to it a (hierarchical) cluster analysis

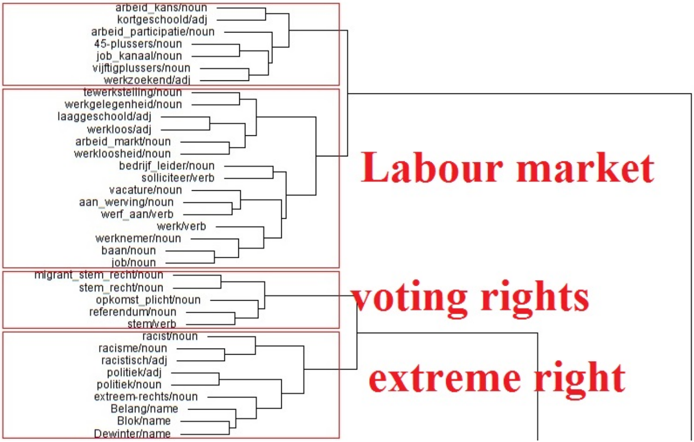


Analysing Semantic Structure

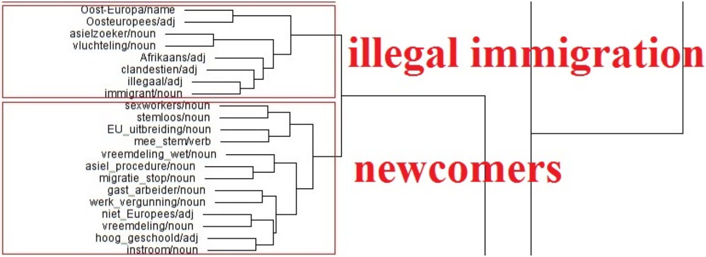
Clusters of contextually related collocations ≈ semantic features
Clusters can be labeled manually



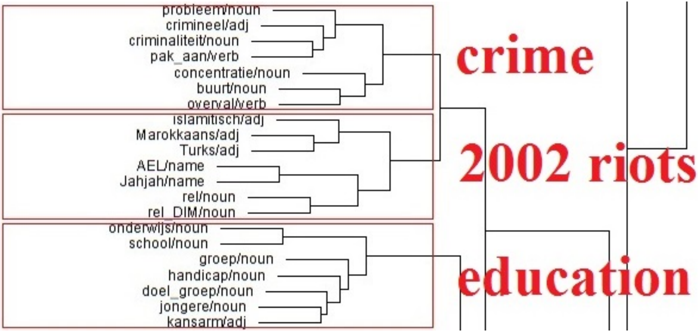
Analysing Semantic Structure



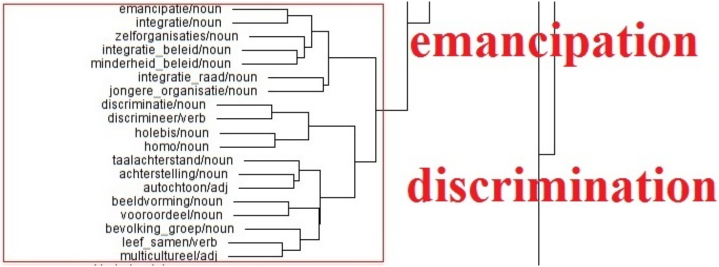
Analysing Semantic Structure



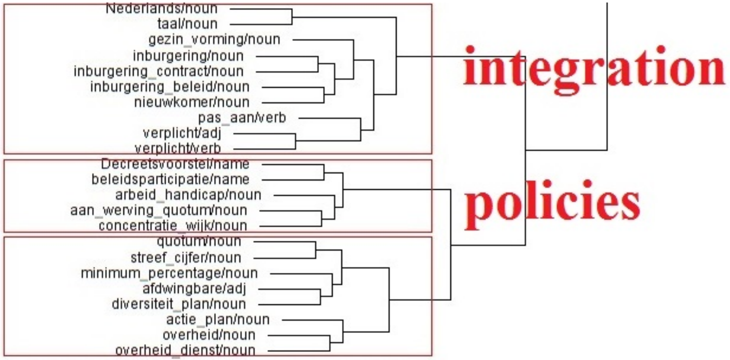
Analysing Semantic Structure



Analysing Semantic Structure

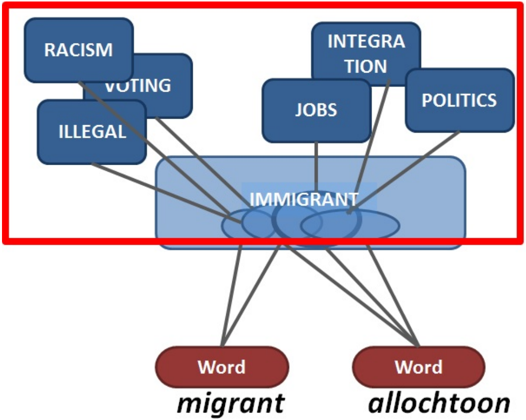


Analysing Semantic Structure



Analysing Semantic Structure

Contextually defined "semantic features" that make up the internal structure of the concept

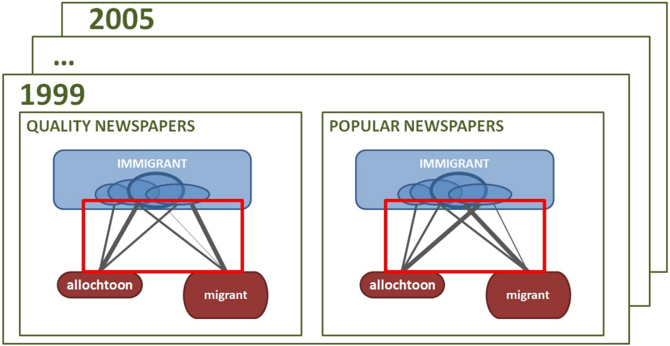


Measuring semantic change in registers

- How strong are *allochtoon* and *migrant* associated with the different context cluster/semantic features
- Is the strength of association the same in quality and popular newspapers?
- Does the strength of association change over time?

Measuring semantic change in registers

What is association strength between semantic features and lexemes in different registers and periods?



Measuring semantic change in registers

STEP 1

Make separate vectors per variant, per year, and per newspaper type

	jobs	racisme	integratie	misdaad	stemrecht	suiker	zon
allochtoon/1999pop	5.3	7.9	6.5	4.0	0.8	0.6	0.0
migrant/1999pop	4.3	8.1	5.7	3.2	6.2	0.5	0.0
allochtoon/1999qual	4.3	2.9	7.5	8.1	0.3	1.6	0.3
migrant/1999qual	4.3	4.2	5.7	3.2	6.2	0.5	0.0
allochtoon/2000pop	5.8	3.5	6.5	5.1	1.3	0.0	0.1
migrant/2000pop	2.9	2.4	4.7	2.2	4.2	0.3	0.7

Measuring semantic change in registers

STEP 2

Make vector per context cluster through aggregation

	jobs	*racisme*	*integratie*	*misdaad*	*stemrecht*	*suiker*	*zon*
jobs	5.3	7.9	6.5	4.0	0.8	0.6	0.0
werk	4.3	8.1	5.7	3.2	6.2	0.5	0.0
arbeidsmarkt	5.3	7.9	6.5	6.0	0.8	0.6	0.1
LABOURMARKET	5.3	7.1	7.7	2.2	6.2	0.4	0.0

Measuring semantic change in registers

STEP 3

Combine variant/year/type vectors and context cluster vectors in 1 matrix

	jobs	racisme	integratie	misdaad	stemrecht	suiker	zon
allochtoon/1999pop	5.3	7.9	6.5	4.0	0.8	0.6	0.0
migrant/1999pop	4.3	8.1	5.7	3.2	6.2	0.5	0.0
allochtoon/1999qual	4.3	2.9	7.5	8.1	0.3	1.6	0.3
migrant/1999qual	4.3	4.2	5.7	3.2	6.2	0.5	0.0
allochtoon/2000pop	5.8	3.5	6.5	5.1	1.3	0.0	0.1
...	...	...	...	...	...	...	...
LABOURMARKET	5.3	7.1	7.7	2.2	6.2	0.4	0.0
...	...	...	...	...	...	...	...

Measuring semantic change in registers

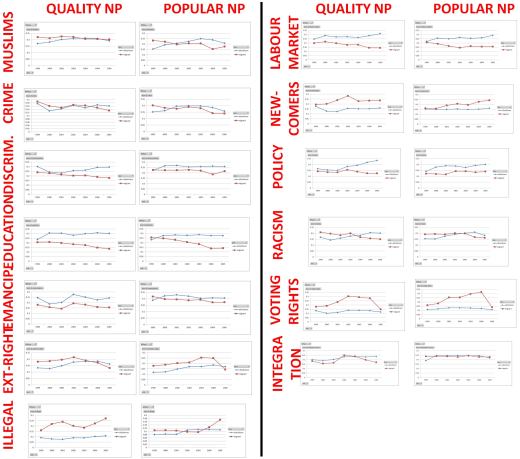
STEP 4
Calculate the cosine similarity (≈ association strength) of each variant/year/type vector to each context cluster vector

	LABOUR	ILLEGAL	EXTREME	POLICY	CRIME	VOTING	RACISM
allochtoon/1999pop	0.3	0.9	0.5	0.0	0.8	0.6	0.0
migrant/1999pop	0.3	0.1	0.7	0.2	0.2	0.5	0.0
allochtoon/1999qual	0.3	0.9	0.5	0.1	0.3	0.6	0.3
migrant/1999qual	0.3	0.2	0.7	0.2	0.2	0.5	0.0
allochtoon/2000pop	0.8	0.5	0.5	0.1	0.3	0.0	0.1
migrant/2000pop	0.9	0.4	0.7	0.2	0.2	0.3	0.7

Measuring semantic change in registers

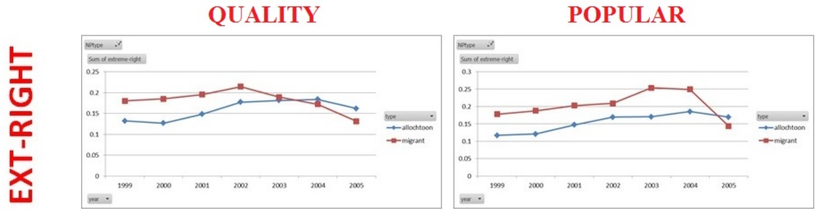
STEP 5

Plot the change of association strength per context cluster and newspaper type



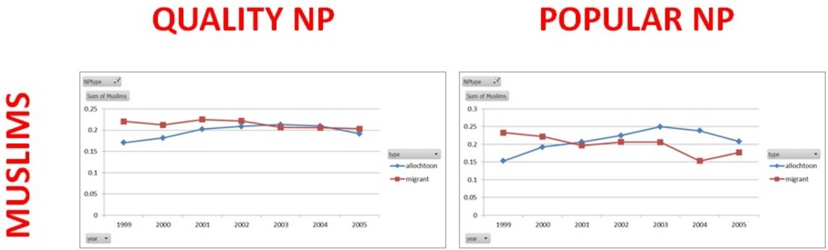
Measuring semantic change in registers

ALLOCHTOON TAKES OVER CONTEXTS FROM MIGRANT



Measuring semantic change in registers

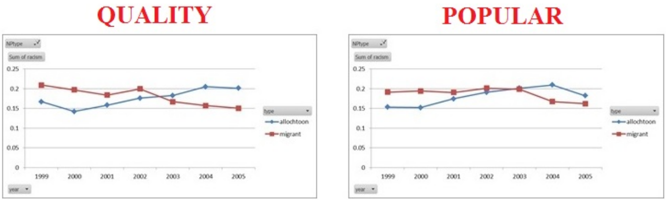
ALLOCHTOON TAKES OVER CONTEXTS FROM MIGRANT



Measuring semantic change in registers

ALLOCHTOON TAKES OVER CONTEXTS FROM MIGRANT

RACISM

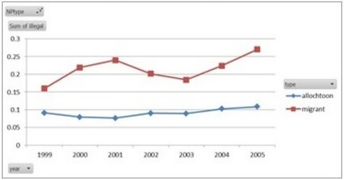


Measuring semantic change in registers

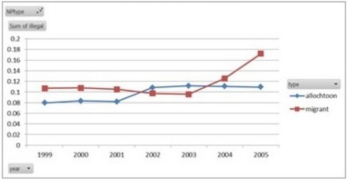
MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON

ILLEGAL

QUALITY

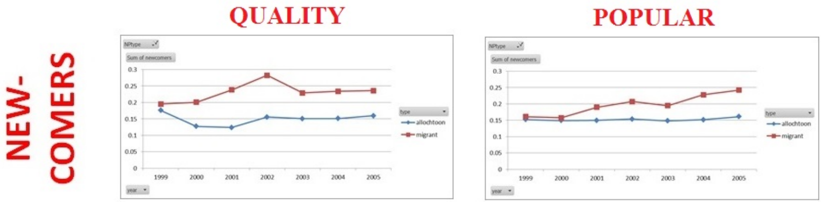


POPULAR



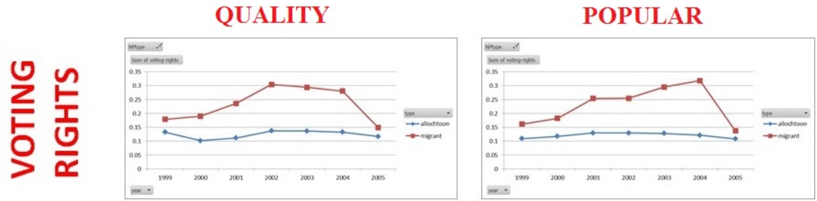
Measuring semantic change in registers

MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON



Measuring semantic change in registers

MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON

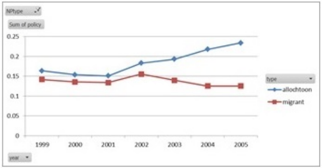


Measuring semantic change in registers

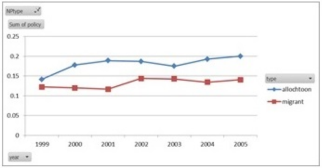
ALLOCHTOON SPECIALIZES RELATIVE TO MIGRANT

POLICY

QUALITY



POPULAR



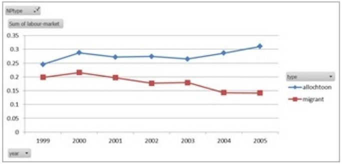
Measuring semantic change in registers

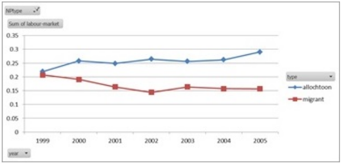
ALLOCHTOON SPECIALIZES RELATIVE TO MIGRANT

QUALITY NP

POPULAR NP

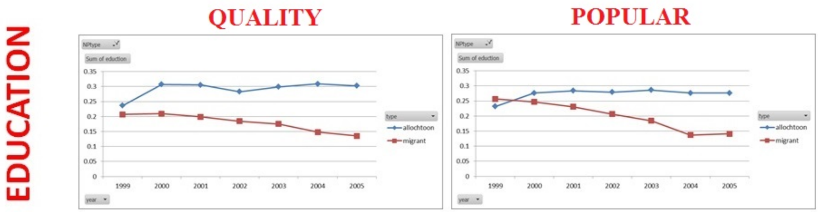
LABOUR MARKET





Measuring semantic change in registers

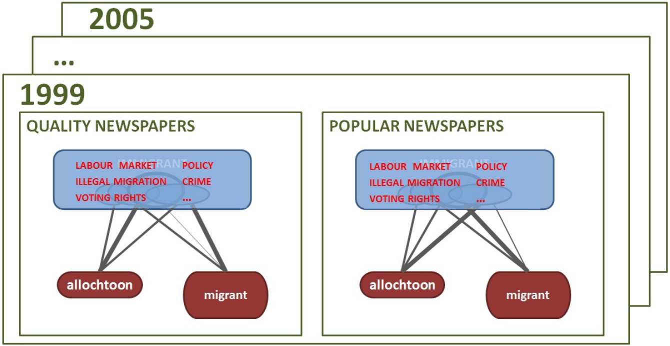
ALLOCHTOON SPECIALIZES RELATIVE TO MIGRANT



Measuring semantic change in registers

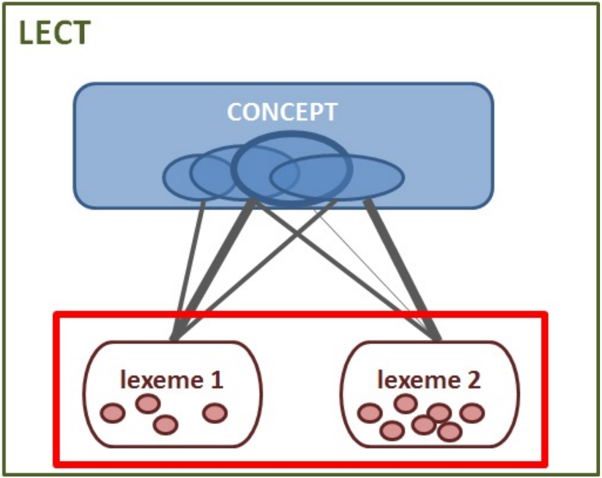
Measuring semantic change in registers

Association strength between semantic features and lexemes differ between registers and changes over time.



Lexical variation on the attestation level

How are the individual occurrences of *allochtoon* and *migrant* distributed over context clusters?



Lexical variation on the attestation level

Make a vector for each attestation of *allochtoon* and *migrant*

op de	arbeidsmarkt	zijn er voor	allochtonen	nauwelijks	jobs
on the	*labour market*	*there are for*	*immigrants*	*scarcely*	*jobs*

Lexical variation on the attestation level

Make a vector for each attestation of *allochtoon* and *migrant*
STEP 1: retrieve the type vectors for each informative context word

	3.2				7.1
	5.1				0.1
	0.2				0.3
	3.1				4.1
	4.7				3.1
	2.2				3.8
op de	arbeidsmarkt	zijn er voor	allochtonen	nauwelijks	jobs
on the	*labour market*	*there are for*	*immigrants*	*scarcely*	*jobs*

Lexical variation on the attestation level

Make a vector for each attestation of *allochtoon* and *migrant*
STEP 2: average over the type vectors of context words

			AVERAGE
3.2		7.1	5.2
5.1		0.1	3.1
0.2		0.3	0.2
3.1		4.1	3.7
4.7		3.1	3.9
2.2		3.8	2.9
arbeidsmarkt	allochtonen	jobs	
labour market	*immigrants*	*jobs*	

Lexical variation on the attestation level

Make a vector for each attestation of *allochtoon* and *migrant*
STEP 3: matrix of exemplar vector with *2nd order* co-occurrences

	jobs	*racisme*	*integratie*	*misdaad*	*stemrecht*	*suiker*	*zon*
allochtoon$_1$	5.3	7.9	6.5	4.0	0.8	0.6	0.0
allochtoon$_2$	4.3	8.1	5.7	3.2	6.2	0.5	0.0
allochtoon$_3$	4.3	2.9	7.5	8.1	0.3	1.6	0.3
migrant$_1$	4.3	4.2	5.7	3.2	6.2	0.5	0.0
migrant$_2$	5.8	3.5	6.5	5.1	1.3	0.0	0.1
migrant$_3$	2.9	2.4	4.7	2.2	4.2	0.3	0.7

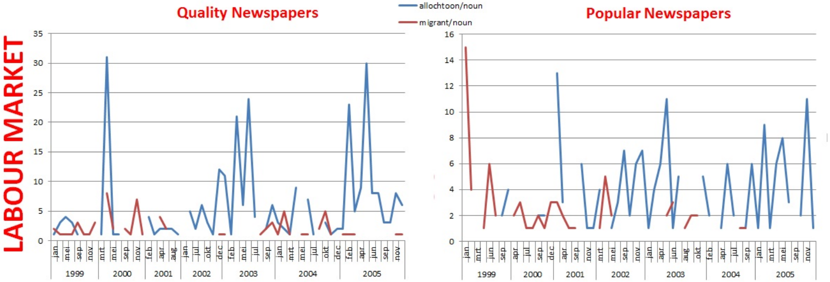
Lexical variation on the attestation level

Make a vector for each attestation of *allochtoon* and *migrant*
STEP 4: calculate similarity matrix between attestation and cluster vectors

	LABOUR	ILLEGAL	EXTREME	POLICY	CRIME	VOTING	RACISM
$allochtoon_1$	0.1	0.9	0.5	0.4	0.8	0.6	...
$allochtoon_2$	0.4	0.3	0.7	0.2	0.2	0.5	...
$allochtoon_3$	0.3	0.9	0.4	0.3	0.3	0.6	...
$migrant_1$	0.3	0.2	0.7	0.3	0.2	0.4	...
$migrant_2$	0.8	0.5	0.5	0.1	0.1	0.0	...
$migrant_3$	0.9	0.4	0.7	0.2	0.2	0.7	...

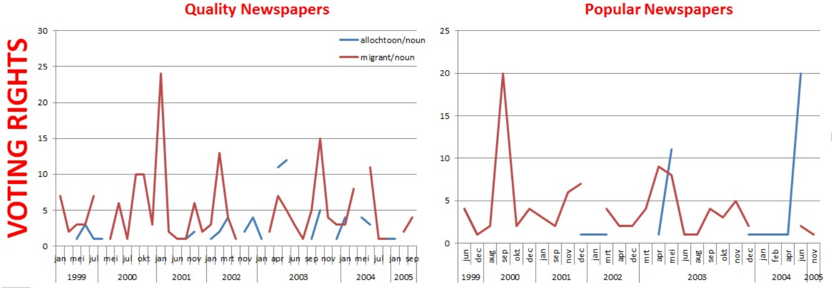
Lexical variation on the attestation level

Same evolution as on aggregated type-level, but with peaks visible



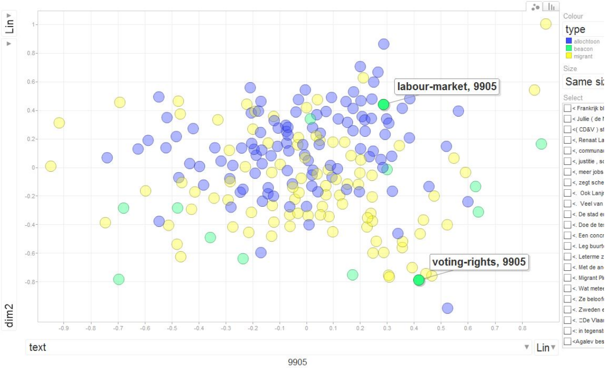
Lexical variation on the attestation level

Same evolution as on aggregated type-level, but with peaks visible



Visualising concordances

Calculate similarity between all tokens
use MDS and googlevis to plot in 2D



Overview

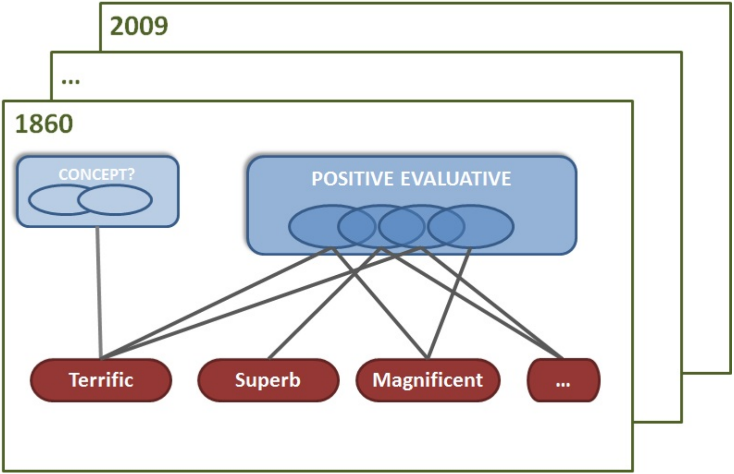
1. Concepts in Cognitive Sociolinguistics
2. Distributional Corpus Analysis
3. Case Study 1: Discourse about immigrants
4. Case Study 2: Positive evaluative adjectives
5. Conclusion

Case study: positive evaluative adjectives

brilliant	cool
delightful	excellent
fabulous	fantastic
good	great
impressive	lovely
magnificent	marvelous
perfect	splendid
superb	terrific
wonderful	

Table: positive evaluative adjectives

Case study: : positive evaluative adjectives



Case study

Corpus

- Corpus of Historical American English (COHA, Davies 2012)
- Period from 1810 to 2009, 400M words, POS-tagged.

Concept: Positive evaluative adjectives

- 1 vector per adjective, per decade (1860-2009)
- modelled by window of 5 words left & right
- 5000 most frequent context words (minus top 100)
- PMI-weighting, cosine similarity

Visualisation

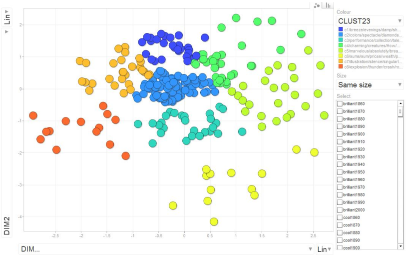
HighD to 2D

- word-decade by context matrix is high dimensional
- first aim is NOT to find latent structure (as with LSA/LDA) but general picture of distributional semantic structuring
- faithful rendering of similarity matrix in 2D: Kruskal's non-metric Multidimensional Scaling
- interpret dimensions with context-labeled clusters

Dynamic and interactive chart

- Motion Charts from Google Chart Tools
- panchronic view to interpret semantic space
- diachronic view to see meaning changes.

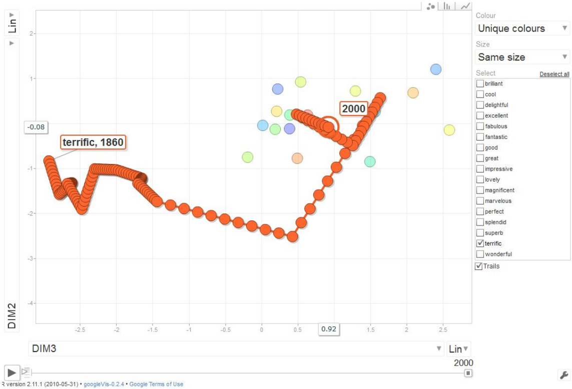
panchronic view for interpretation of semantic space



Clusters with most typical contextwords of adjectives:

- cluster 2 (centre, light blue): positive evaluated things (*colors, spectacle, performance*) ⇒ centre of the plot, expressing the core meaning of the adjectives
- cluster 8 (red, lower left): loud and frightening things (*explosion, thunder, crash*) ⇒ periphery of the plot, expressing non-related meaning

diachronic motion chart to see meaning change



Trajectory of *terrific* from 1860 to 2000, moving from the peripheral cluster of "frightening things" to the central cluster of "positive evaluated things", indicative of its meaning change

For more information:

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

kris.heylen@kuleuven.be

Acknowledgements:

Dirk Geeraerts

Dirk Speelman

Thomas Wielfaert

Martin Hilpert

Georgiana Dinu