The Sem·metrix Project: Scaling up the Profile-Based Measurement of Lexical Variation

Kris Heylen & Yves Peirsman

KULeuven
Quantitative Lexicology and Variational Linguistics

Aim of the Sem·metrix Project

- A Corpus-based method to study lexical variation that starts from the actual lexical options available to express a concept
  - developed by Geeraerts, Speelman & Grondelaers 1999
  - avoids thematic bias and controls for polysemy
  - differences between Belgian and Netherlandic Dutch
- Automatizing this method by exploiting advanced computational linguistic techiques
  - lexical variation research on a large scale (the whole lexicon)
  - language independent, highly portable

Overview

1. Profile-based Measurement of Lexical Variation
2. Project build-up
3. Generating Synonyms
4. First results
5. Conclusions

Profile-based Measurement of Lexical Variation

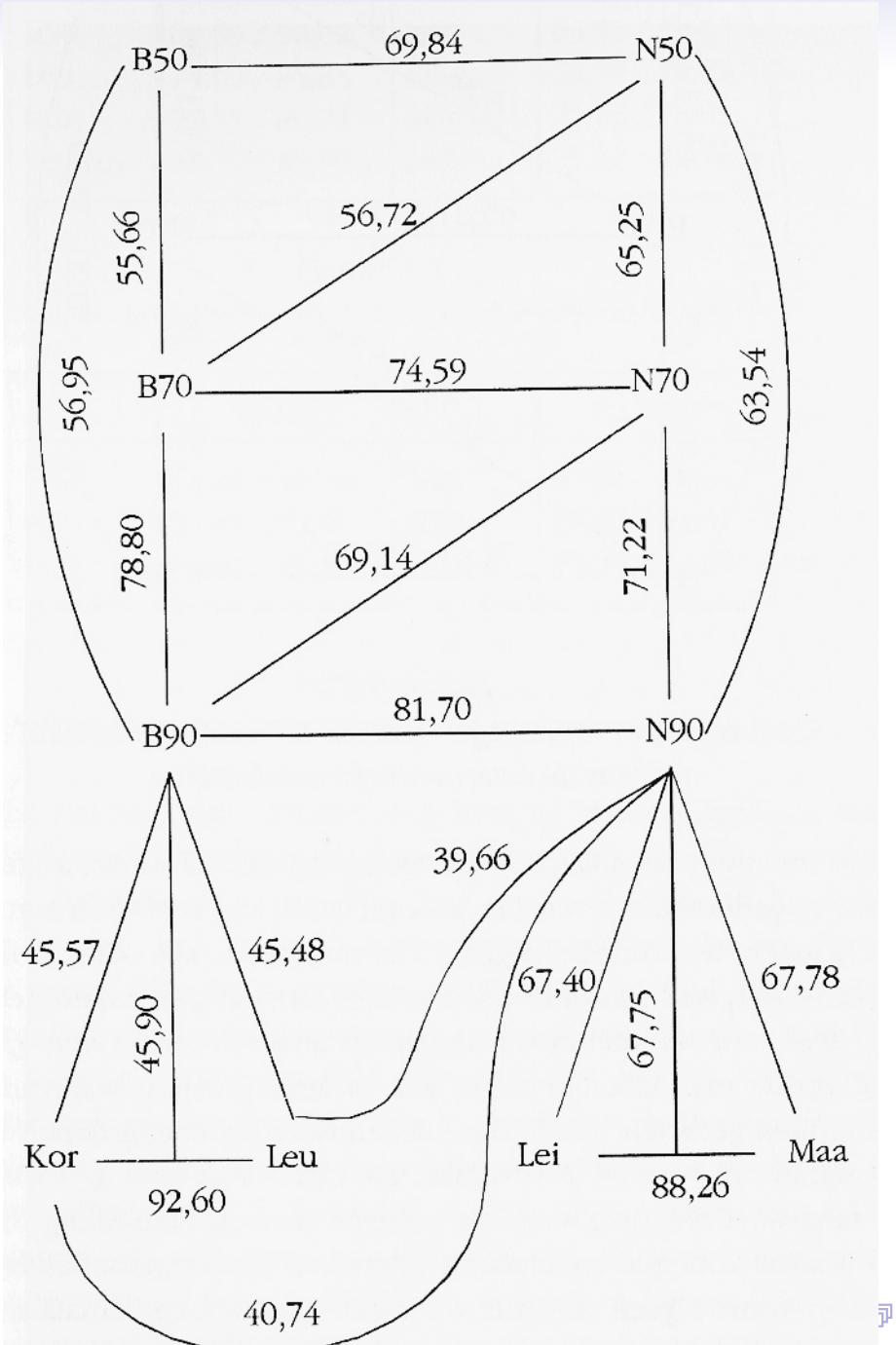
Onomasiological perspective: Do different varieties use different lexemes to express a specific concept?

- Define a set of synonoyms = profile
- collect all instances from 2 corpora (B vs NL) + disambiguate
- Compute relative frequency of synonyms in 2 corpora
- Overlap in relative frequency = uniformity measure

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

Profile-based Measurement of Lexical Variation

- Extend the appoach to multiple profiles for general assesment
- average over profiles
- possibly weighted for concept frequency
- Geeraerts, Speelman & Grondelaers 1999, 2003 for Dutch
  - **Semantic fields:** clothing and football terms
  - **Region:** Belgium vs Netherlands
  - **Register:** newspaper vs shop windows
  - **Diachronic:** 50's, 70's, 90's


B50
N50
69,84
56,72
55,66
65,25
56,95
63,54
B70
74,59
N70
78,80
69,14
71,22
B90
81,70
N90
39,66
45,57
45,48
67,40
67,78
45,90
67,75
Kor
Leu
Lei
Maa
92,60
88,26
40,74

Profile-based Measurement of Lexical Variation

Advantage

- Corpus-based, empirical and quantified
- Onomasiological: actual lexical choices faced by speakers
- avoids thematic bias and controls for polysemy

Problems

- Time consuming manual definition of profiles
- Time consuming disambiguation of polysemous lexemes
- not readily scalable to many profiles or other languages

$\Rightarrow$ the sem· metrix project

Overview

1. Profile-based Measurement of Lexical Variation
2. Project build-up
3. Generating Synonyms
4. First results
5. Conclusions

Project build-up: Aims

Short term: Automazing profile generation and disambiguation

- using existing NLP techniques: Synonymy Extraction and Word Sense Disambiguation
- exploit application oriented computational linguistics for variationist research
- development of socio-lectometric tools for large scale lexical variation research

Long term: In depth study of lexical variation in Dutch

- investigate a large number of profiles to obtain overall assesment
- extend profile-based approach to non lexical items

Project build-up

1. Identifying profile concepts
   - key-words method to extract frequent concepts
2. Finding synonymous lexemes for the profile concepts
   - Distributional similarity over contexts
3. Disambiguation of lexemes in context
   - occurrences are only relevant when they refer to profile concept (jeans = trousers $\neq$ cloth)
4. Delineation of language varieties
   - instead of top-down (informal vs formal) bottom-up
   - clustering subcorpora on the basis of morpho-syntactic features (Biber 1999)

Overview

1. Profile-based Measurement of Lexical Variation
2. Project build-up
3. Generating Synonyms
4. First results
5. Conclusions

Generating Synonyms

Basic principle: Words that occur in similar contexts will have similar meanings

. . . He wore baggy **trousers** *under a white shirt . . .*
. . . They live in a brick **house** *with a porch. . . .*
. . . The more you wash your **jeans** *the tighter it will fit . . .*
. . . She wore comfortable linen **slacks** *and a red blouse . . .*

$\Rightarrow$ context distributional similarity $\approx$ semantic similarity

What context features should be taken into account?

Collocates

- window of e.g. 5 words left and right
- looser associative semantic relations (Kilgarrif & Yallop 2000)

Syntactic dependency relations

- subject, object, prepositional complement,...
- a parsed corpus is needed
- tighter, synonym-like semantic relations

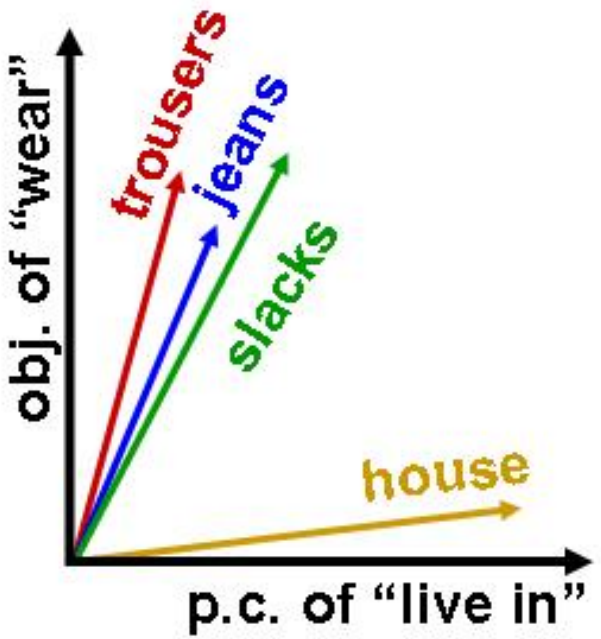
Generating Synonyms

context features are extracted from a corpus for each target word (e.g. all nouns) and put into a vector

	obj. of wear	obj. of wash	subj. of shrink	p.c. of live in	...
slacks	19	15	14	0	...
jeans	78	43	39	1	...
trousers	56	27	33	0	...
house	0	1	0	78	...
...	...	...	...	...	...

Generating Synonyms

weighted vectors are mapped into geometrical space



distance between vectors is calculated $\Rightarrow$ semantic distance (the inverse is a similarity measure)

Generating Synonyms

results in a word by word similarity matrix:

	slacks	jeans	trousers	house	...
slacks	1	.95	.91	.08	...
jeans	.95	1	.89	.05	...
trousers	.91	.89	1	.03	...
house	.08	.05	.03	1	...
...	...	...	...	...	...

Words with high mutual similarity are clustered into 'profiles'

Overview

1. Profile-based Measurement of Lexical Variation
2. Project build-up
3. Generating Synonyms
4. First results
5. Conclusions

First results

Data

- Twente NieuwsCorpus: 300M (12 y. of Dutch newspapers)
- Automatically parsed with Alpino

Context features

- collocates (5 words L+R)
- syntactic dependency features (8: subject, object, prepositional complement, advervbial PP, adjective, postmodifying PP, apposition ,conjunction)

Evaluation

- clothing and football profiles
- 1000 word random sample against WordNet Dutch

First results: Football terms

Collocates

goal	*doelpunt, treffer, foutje, fout, keuze*
strafschop	*penalty, strafworp, strafbal, invaller, thuisclub*
hoekschop	*corner, trap, invaller, voorzet, openingstreffer*

Syntactic dependency features

goal	*doelpunt, treffer, gelijkmaker, openingstreffer, strafschop*
strafschop	*penalty, doelpunt, treffer, strafbal, gelijkmaker*
hoekschop	*corner, voorzet, trap, pass, schot*

First results: 1000 word sample

- random sample of 1000 nouns from the corpus
- relations found among 10 most related words were checked against relations in EuroWordNet Dutch

	syn.	hypo.	hyper.	cohyp.	all 4
syntax	6.3	4.0	4.2	17.0	31.5
bag-of-words	4.2	2.7	2.8	12.2	21.9

$\Rightarrow$ narrow coverage of EWN: problematic evaluation standard

TARGET	1	2	3	4	5
Zweeds	Maleis	Italiaans	ServoKroatisch	Duits	Japans
afgrijzen	afschuw	verbazing	verbijstering	ontzetting	verwondering
bomaanslag	aanslag	zelfmoordaanslag	bomexplosie	moordaanslag	zelfmoordactie
competitiewedstrijd	competitieduel	thuiswedstrijd	uitwedstrijd	wedstrijd	bekerfinale
alcoholisme	drugverslaving	drankmisbruik	incest	drugmisbruik	drugsgebruik
nier	lever	milt	alvleesklier	long	darm
aardbeving	beving	aardschok	overstroming	bosbrand	vulkaanuitbarsting
koersval	koersdaling	koersstijging	waarddaling	daling	waardeverminderi
oestrogeen	cortisol	testosteron	progesteron	hormoon	statines
oester	kreeft	mossel	tarbot	asperge	garnaal
incest	kindermishandeling	verkrachting	ontucht	sodomie	overspel
obstipatie	winderigheid	verstopping	diarree	nierziekte	hartkwaal
letsel	verwonding	rookvergiftiging	hoofdletsel	snijdwonde	schedelbasisfract
straaljager	gevechtsvliegtuig	F-16	jachtvliegtuig	bommenwerper	gevechttoestel
gelach	boegeroep	gejuich	lachsalvo	gejoel	hoongelach
cyanide	blauwzuurgas	arseen	cadmium	arsenicum	styreen
verslagenheid	vertwijfeling	ontreddering	verbijstering	radeloosheid	ongeloof
country	folk	bluegrass	gospel	blues	reggae
arbeidskosten	loonkosten	energiekost	pensioenkost	personeelskosten	detailhandelomzet
schurft	syfilis	hivaids	tuberculose	malaria	tbc
toewijding	wilskracht	ijver	overgave	volharding	doorzettingsvermo

Overview

1. Profile-based Measurement of Lexical Variation
2. Project build-up
3. Generating Synonyms
4. First results
5. Conclusions

Conclusions

- Established profile-based approach of lexical variation.
- Sem·metrix project: automatisation to scale up the approach
- First step: automatic generation of profiles
- First results are promising, but room for improvement
- Evaluation has to be refined

For more information:

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

kris.heylen@arts.kuleuven.be

yves.peirsman@arts.kuleuven.be