



The Sem·matrix Project: Towards the Large-Scale Measurement of Lexical Variation

Yves Peirsman & Kris Heylen



KU Leuven

Quantitative Lexicology and Variational Linguistics

Aim of the Sem-matrix 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, 2003
 - avoids **thematic bias** and controls for **polysemy**
 - differences between Belgian and Netherlandic Dutch
- **Automatizing** this method by exploiting advanced computational linguistic techniques
 - lexical variation research on a **large scale** (the whole lexicon)
 - **language independent**, highly portable sociolectometric tool



Overview

1. Profile-based Measurement of Lexical Variation
2. Automatic Synonymy Retrieval
3. First results
4. Example Case Study
5. Conclusions



1. Profile-based Measurement of Lexical Variation

Corpus-based study of word use in language varieties

Do different varieties use different lexemes to express a specific concept?

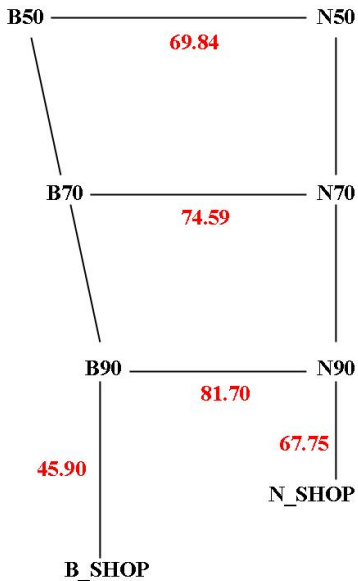
- Define a set of synonyms = profile
- collect all instances from 2 corpora (B vs NL) + disambiguate
- Compute relative frequency of synonyms in 2 corpora
- Overlap in relative frequency = uniformity measure

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



1. Profile-based Measurement of Lexical Variation

- Extend the approach 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:** 50s, 70s, 90s



1. Profile-based Measurement of Lexical Variation

Advantages

- 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

⇒ sem·metrix project



2. Automatic Synonymy Retrieval

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

*... He wore baggy **trousers** under a comfortable white shirt ...*

*... They live in a brick **house** with a porch. ...*

*... The more you wash your **jeans** the tighter they will fit ...*

*... The comfortable linen **slacks** she wore didn't really fit ...*

Context features: bag-of-words or syntactic dependency relations



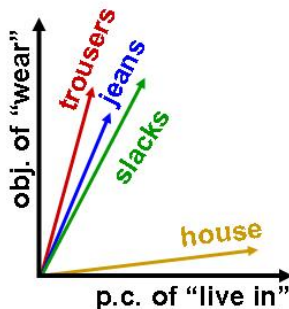
2. Automatic Synonymy Retrieval

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	...
...

2. Automatic Synonymy Retrieval

weighted vectors are mapped into geometrical space



distance between vectors \Rightarrow semantic distance

3. First results

Data

- Twente Nieuws Corpus: 300 million words (4 years material of 5 Dutch newspapers)
- Parsed and lemmatized

Context features

- collocates (5 words L+R)
- syntactic dependency features



3. First results

Collocational features

goal	<i>doelpunt, treffer, foutje, fout, keuze</i>
strafschop	<i>penalty, strafworp, strafbal, invaller, thuisclub</i>
hoekschop	<i>corner, trap, invaller, voorzet, openingstreffer</i>

Syntactic dependency features

goal	<i>doelpunt, treffer, gelijkmaker, openingstreffer, strafschop</i>
strafschop	<i>penalty, doelpunt, treffer, strafbal, gelijkmaker</i>
hoekschop	<i>corner, voorzet, trap, pass, schot</i>



3. First results

Experiments

- random sample of 1,000 nouns from the corpus
- Six different context models
- relations found among 10 most related words were checked against relations in EuroWordNet Dutch

	syn.	hypo.	hyper.	cohyp.	all 4	in EWN
syn	6.3	4.0	4.2	16.9	31.4	6215
coll, 5w	4.2	2.7	2.8	12.2	21.8	5645
syn, RI	2.1	1.3	1.9	8.6	13.9	7275
coll, 5w, 2000	2.5	1.8	1.7	9.8	15.9	5739
coll, 5w, RI	2.4	1.3	2.1	8.3	14.2	5598
coll, 50w, 2000	1.6	0.9	1.1	6.2	9.8	6682

Finding Syntactic context models work best.

Future work Experiments with different parameter settings



Sample of results using syntactic features

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	alveesklier	long	darm
aardbeving	beving	aardschok	overstroming	bosbrand	vulkaanuitbarsting
koersval	koersdaling	koersstijging	waarddaling	daling	waardevermindering
oestrogeen	cortisol	testosteron	progesteron	hormoon	statines
oester	kreeft	mossel	tarbot	asperge	garnaal
incest	kinder mishandeling	verkrachting	ontucht	sodomie	overspel
obstipatie	winderigheid	verstopping	diarree	nierziekte	hartkwaal
letsel	verwonding	rookvergiftiging	hoofdletsel	snijdwonde	schedelbasisfractuur
straaljager	gevechtsvliegtuig	F-16	jachtvliegtuig	bommenwerper	gevechtstoestel
gelach	boegeroep	gejuich	lachsalvo	gejoel	hoongelach
cyanide	blauwzuurgas	arsen	cadmium	arsenicum	styreen
verslagenheid	vertwijfeling	ontredning	verbijstering	radeloosheid	ongeloof
country	folk	bluegrass	gospel	blues	reggae
arbeidskosten	loonkosten	energiekost	pensioenkost	personeelskosten	detailhandelomzet
schurft	syfilis	hiv aids	tuberculose	malaria	tbc
toewijding	wilskracht	ijver	overgave	volharding	doorzettingsvermogen

4. Example Case Study

Research Question

Do the main Dutch newspapers differ in their use of English words?

Test words

keeper

corner

penalty

goal

setpoint



4. Example Case Study

Research Question

Do the main Dutch newspapers differ in their use of English words?

Test words

keeper	doelman
corner	hoekschop
penalty	strafschop
goal	doelpunt, treffer
setpoint	setpunt

4. Example Case Study

SETPUNT	nrc	volkskrant	parool	trouw	ad
setpoint	45	26	14	14	9
setpunt	10	23	12	15	21
DOELMAN					
keeper	542	776	1355	520	1067
doelman	1598	2732	1580	1366	3170
HOEKSCHOP					
corner	135	197	272	150	204
hoekschop	99	259	125	101	176
STRAFSCHOP					
penalty	260	379	463	150	506
strafschop	1040	1423	1105	707	1512
DOELPUNT					
goal	314	613	820	491	1091
treffer	1258	2239	1320	1324	2829
doelpunt	2540	4249	4014	1838	4336



4. Example Case Study

SETPUNT	nrc	volkskrant	parool	trouw	ad
setpoint	.82	.53	.54	.48	.30
setpunt	.18	.47	.46	.52	.70
DOELMAN					
keeper	.25	.22	.46	.28	.25
doelman	.75	.78	.54	.72	.75
HOEKSCHOP					
corner	.58	.43	.69	.60	.54
hoekschop	.42	.57	.31	.40	.46
STRAFSCHOP					
penalty	.20	.21	.30	.18	.25
strafschop	.80	.79	.70	.82	.75
DOELPUNT					
goal	.07	.08	.13	.13	.13
treffer	.31	.32	.21	.36	.34
doelpunt	.62	.60	.65	.50	.53



4. Example Case Study

SETPUNT	nrc	volkskrant	parool	trouw	ad
setpoint	.82	.53	.54	.48	.30
setpunt	.18	.47	.46	.52	.70
DOELMAN					
keeper	.25	.22	.46	.28	.25
doelman	.75	.78	.54	.72	.75
HOEKSCHOP					
corner	.58	.43	.69	.60	.54
hoekschop	.42	.57	.31	.40	.46
STRAFSCHOP					
penalty	.20	.21	.30	.18	.25
strafschop	.80	.79	.70	.82	.75
DOELPUNT					
goal	.07	.08	.13	.13	.13
treffer	.31	.32	.21	.36	.34
doelpunt	.62	.60	.65	.50	.53



4. Example Case Study

Overlap matrix

	nrc	volkskrant	ad	parool	trouw
nrc	1	.97	.94	.87	.93
volkskrant		1	.94	.86	.92
ad			1	.86	.97
parool				1	.85
trouw					1

4. Example Case Study

SETPUNT	nrc	volkskrant	parool	trouw	ad
setpoint	.82	.53	.54	.48	.30
setpunt	.18	.47	.46	.52	.70
DOELMAN					
keeper	.25	.22	.46	.28	.25
doelman	.75	.78	.54	.72	.75
HOEKSCHOP					
corner	.58	.43	.69	.60	.54
hoekschop	.42	.57	.31	.40	.46
STRAFSCHOP					
penalty	.20	.21	.30	.18	.25
strafschop	.80	.79	.70	.82	.75
DOELPUNT					
goal	.07	.08	.13	.13	.13
treffer	.31	.32	.21	.36	.34
doelpunt	.62	.60	.65	.50	.53



4. Example Case Study

Problematic issues

- Automatic cutoff to find profile boundaries
⇒ Alternatively, overlap can be measured wrt broader groups of nouns
- Disambiguation of the nouns in the corpus
⇒ Word Sense Disambiguation techniques



5. Conclusions

- Established profile-based approach of lexical variation.
- Sem-matrix project: automatism 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