



Het sem-matrix-project

De profielgebaseerde meting van lexicale
variatie op een grotere schaal

Kris Heylen & Yves Peirsman



University of Leuven

RU Quantitative Lexicology and Variational Linguistics

Doelstelling van sem-metrix

- Kwantificeren van **verschillen in woordgebruik** tussen variëteiten van het Nederlands (e.g. Belgisch vs. Nederlands Nederlands)
- **Automatisering** van bestaande onderzoeksmethodes door middel van computationeel linguïstische technieken....
- met als sleuteltechniek **semantische afstandsmeting** aan de hand van contextueel-distributioniële similariteit

Overzicht

1. Achtergrond:
 - Profielgebaseerde meting van lexicale variatie
2. sem-matrix en zijn deelprojecten
3. Genereren van profielen (*synsets*)
 - Data en dataverrijking
 - Contextmodel
 - Afstandsmeting en clustering
 - Evaluatie
4. Eerste resultaten
 - Betekenisverwante woorden vinden aan de hand van collocaten en dependentierelaties → Yves Peirsman

Overzicht

1. Achtergrond:

- Profielgebaseerde meting van lexicale variatie

2. sem-matrix en zijn deelprojecten

3. Genereren van profielen (*synsets*)

- Data en dataverrijking
- Contextmodel
- Afstandsmeting en clustering
- Evaluatie

4. Case study

- Hoe geschikt is Wordnet Dutch als evaluatiestandaard?

Achtergrond

Profielgebaseerde meting van
lexicale variatie (1999)



Achtergrond

Profielgebaseerde meting van
lexicale variatie (1999)

- Hoe verschilt het woordgebruik in
verschillende NI. variëteiten?



Achtergrond

Profielgebaseerde meting van
lexicale variatie (1999)

- Hoe verschilt het woordgebruik in verschillende NI. variëteiten?
- In hoeverre worden verschillende lexemen gebruikt om **een zelfde concept** uit te drukken
JEANS (jeans, spijkerbroek)

Convergentie en divergentie in de
Nederlandse woordenschat

Een onderzoek naar kleding- en voetbaltermen

Dirk Geeraerts
Stefan Grondelaers
Dirk Speelman

Achtergrond

Profielgebaseerde meting van
lexicale variatie (1999)

- Hoe verschilt het woordgebruik in verschillende NI. variëteiten?
- In hoeverre worden verschillende lexemen gebruikt om **een zelfde concept** uit te drukken
JEANS (jeans, spijkerbroek)
- Profiel = set van synoniemen voor bepaald concept

Convergentie en divergentie in de
Nederlandse woordenschat

Een onderzoek naar kleding- en voetbaltermen

Dirk Geeraerts
Stefan Grondelaers
Dirk Speelman

Achtergrond

Profielgebaseerde meting van
lexicale variatie (1999)

- Hoe verschilt het woordgebruik in verschillende NI. variëteiten?
- In hoeverre worden verschillende lexemen gebruikt om **een zelfde concept** uit te drukken
JEANS (jeans, spijkerbroek)
- Profiel = set van synoniemen voor bepaald concept
- Uniformiteit van variëteiten = mate waarin de distributie over synoniemen overlapt.

Convergentie en divergentie in de
Nederlandse woordenschat

Een onderzoek naar kleding- en voetbaltermen

Dirk Geeraerts
Stefan Grondelaers
Dirk Speelman

Achtergrond

Voorbeeld: concept STRAFSCHOP

	NL '90	B '90	OVERLAP
elfmeter	0 (0%)	23 (14%)	0%
elfmetertrap	3 (1%)	1 (1%)	1%
penalty	172 (64%)	31 (18%)	18%
strafschop	94 (35%)	115 (68%)	35%
UNIFORMITEIT:			54%

Achtergrond

Voorbeeld: concept STRAFSCHOP

	NL '90	B '90	OVERLAP
elfmeter	0 (0%)	23 (14%)	0%
elfmetertrap	3 (1%)	1 (1%)	1%
penalty	172 (64%)	31 (18%)	18%
strafschop	94 (35%)	115 (68%)	35%
UNIFORMITEIT:			54%

- Voor concepten uit kleding en voetbal
- Gestratifiëerd voor Regio, Register, Periode

Achtergrond

Voorbeeld: concept STRAFSCHOP

	NL '90	B '90	OVERLAP
elfmeter	0 (0%)	23 (14%)	0%
elfmetertrap	3 (1%)	1 (1%)	1%
penalty	172 (64%)	31 (18%)	18%
strafschop	94 (35%)	115 (68%)	35%
UNIFORMITEIT:			54%

- Voor concepten uit kleding en voetbal
- Gestratifiëerd voor Regio, Register, Periode
- Voordeel: vermijding van thematische bias
- Probleem: tijdrovende manuele opstelling van profielen en tokendisambiguering in context

sem-metrix

Doelstelling:

- **Korte termijn:** Automatisering van profielgenerering en disambiguering in context door middel van NLP-technieken
 - > ontwikkeling van “sociolectometrische tools”
- **Lange termijn:** Onderzoek van een groot aantal profielen om een vollediger beeld te krijgen van de lexicale variatie tussen variëteiten van het Nederlands
 - > empirische variatielinguïstiek

sem-matrix

Deelprojecten:

1. Identificeren van concepten

- Relevante concepten die vaak optreden (eerst nomina)
- Keyword-methodes

sem-matrix

Deelprojecten:

1. Identificeren van concepten

- Relevante concepten die vaak optreden (eerst nomina)
- Keyword-methodes

2. Genereren van profielen (*synsets*) voor elk concept

- Set van synoniemen per concept (keywords als seeds)
- Distributionele similariteit van contextfeatures om betekenisverwante woorden (*types*) op te sporen

sem-matrix

Deelprojecten:

1. Identificeren van concepten

- Relevante concepten die vaak optreden (eerst nomina)
- Keyword-methodes

2. Genereren van profielen (*synsets*) voor elk concept

- Set van synoniemen per concept (keywords als seeds)
- Distributionele similariteit van contextfeatures om betekenisverwante woorden (types) op te sporen

3. Disambigueren van tokens in context

- Tokens zijn alleen relevant als ze de conceptbetekenis van het profiel hebben (bv. SPIJKER(spijker, nagel)) -> WSD

sem-matrix

Deelprojecten:

1. Identificeren van concepten

- Relevante concepten die vaak optreden (eerst nomina)
- Keyword-methodes

2. Genereren van profielen (*synsets*) voor elk concept

- Set van synoniemen per concept (keywords als seeds)
- Distributionele similariteit van contextfeatures om betekenisverwante woorden (types) op te sporen

3. Disambigueren van tokens in context

- Tokens zijn alleen relevant als ze de conceptbetekenis van het profiel hebben (bv. SPIJKER(spijker, nagel)) -> WSD

4. Afbakenen van de variëteiten

- ipv top-down (B><NL), bottom-up
- Clusteren van subcorpora op basis van morfo-syntactische kenmerken (Biber 1999)

sem-matrix

Deelprojecten:

1. Identificeren van concepten

- Relevante concepten die vaak optreden (eerst nomina)
- Keyword-methodes

2. Genereren van profielen (*synsets*) voor elk concept

- Set van synoniemen per concept (keywords als seeds)
- Distributionele similariteit van contextfeatures om betekenisverwante woorden (types) op te sporen

3. Disambigueren van tokens in context

- Tokens zijn alleen relevant als ze de conceptbetekenis van het profiel hebben (bv. SPIJKER(spijker, nagel)) -> WSD

4. Afbakenen van de variëteiten

- ipv top-down (B><NL), bottom-up
- Clusteren van subcorpora op basis van morfo-syntactische kenmerken (Biber 1999)

Overzicht

1. Achtergrond:
 - Profielgebaseerde meting van lexicale variatie
2. sem-matrix en zijn deelprojecten
3. Genereren van profielen (*synsets*)
 - Data en dataverrijking
 - Contextmodel
 - Afstandsmeting en clustering
 - Evaluatie

Genereren van profielen (*synsets*)

Data:

- eerste fase: krantenmateriaal
 - “Makkelijk” in grote hoeveelheden verkrijgbaar
 - Grotere controle over extralinguïstische kenmerken
 - Minst problematisch voor NLP-tools

Genereren van profielen (*synsets*)

Data:

- eerste fase: krantenmateriaal
 - “Makkelijk” in grote hoeveelheden verkrijgbaar
 - Grotere controle over extralinguïstische kenmerken
 - Minst problematisch voor NLP-tools
- Reeds verkregen:
 - Condiv corpus: 20M (15M B / 5M NL)
 - Condiv+: 900M, maar copyrightproblemen en vooral NL
 - Twente Nieuwscorpus: 78M (NL)
 - Nederlandse Wikipedia

Genereren van profielen (*synsets*)

Data:

- eerste fase: krantenmateriaal
 - “Makkelijk” in grote hoeveelheden verkrijgbaar
 - Grotere controle over extralinguïstische kenmerken
 - Minst problematisch voor NLP-tools
- Reeds verkregen:
 - Condiv corpus: 20M (15M B / 5M NL)
 - Condiv+: 900M, maar copyrightproblemen en vooral NL
 - Twente Nieuwscorpus: 300M (NL) (deels overlappend met C+)
 - Nederlandse Wikipedia
- Extra Belgisch materiaal:
 - Onderhandelingen met Roularta (Knack)
 - Onderhandelingen met Mediargus (Belgische kranten)

Genereren van profielen (*synsets*)

Data-preprocessing:

- Lemmatiseren / taggen
 - Synoniemen op lemma-niveau, gedisambigued voor POS
 - WOTAN tagger-lemmatizer

Genereren van profielen (*synsets*)

Data-preprocessing:

- Lemmatiseren / taggen
 - Synoniemen op lemma-niveau, gedisambigueerd voor POS
 - WOTAN tagger-lemmatizer
- Parseren:
 - Dependenterelaties als mogelijke contextfeatures
 - ALPINO-parser

Genereren van profielen (*synsets*)

Data-preprocessing:

- Lemmatiseren / taggen
 - Synoniemen op lemma-niveau, gedisambigueerd voor POS
 - WOTAN tagger-lemmatizer
- Parseren:
 - Dependenterelaties als mogelijke contextfeatures
 - ALPINO-parser
- Infrastructuur:
 - 2 processor HP-server (3.4GHz processor) met 4GB RAM
 - 147GB eigen schijfruimte, 2TB externe schijven.
 - Redhat Enterprise Linux 4.

Genereren van profielen (*synsets*)

Contextmodel:

- Betekenisverwante woorden komen in gelijkaardige contexten voor -> distributionele similariteit

Genereren van profielen (*synsets*)

Contextmodel:

- Betekenisverwante woorden komen in gelijkaardige contexten voor -> distributionele similariteit
- Contextmodel: welke contextfeatures hebben synoniemen gemeenschappelijk?

Genereren van profielen (*synsets*)

Contextmodel:

- Betekenisverwante woorden komen in gelijkaardige contexten voor -> distributionele similariteit
- Contextmodel: welke contextfeatures hebben synoniemen gemeenschappelijk? hypothese:
 - ongestructureerde coöccurentie -> losse verwantschap
 - gedeelde dependentierelaties -> enge verwantschap

Genereren van profielen (*synsets*)

Contextmodel:

- Betekenisverwante woorden komen in gelijkaardige contexten voor -> distributionele similariteit
- Contextmodel: welke contextfeatures hebben synoniemen gemeenschappelijk? hypothese:
 - ongestructureerde coöccurentie -> losse verwantschap
 - gedeelde dependentierelaties -> enge verwantschap
- Verschillende contextmodellen:
 - Bag-of-words technieken: LSA, collocaten (1^{ste}/2^{de} orde)
 - Syntactische features: dependentierelaties van ALPINO

Genereren van profielen (*synsets*)

Contextmodel:

- Betekenisverwante woorden komen in gelijkaardige contexten voor -> distributionele similariteit
- Contextmodel: welke contextfeatures hebben synoniemen gemeenschappelijk? hypothese:
 - ongestructureerde coöccurentie -> losse verwantschap
 - gedeelde dependentierelaties -> enge verwantschap
- Verschillende contextmodellen:
 - Bag-of-words technieken: LSA, collocaten (1^{ste}/2^{de} orde)
 - Syntactische features: dependentierelaties van ALPINO
- Selectie van features op basis van
 - Indicaties uit de literatuur (Van der Plas & Bouma 2004)
 - Modelering van contextfeatures in functie van sem. relaties

Genereren van profielen (*synsets*)

Afstandsmeting:

- Matrix met voor elke lemma, vector van contextfeatures
- Eventueel Singuliere waardenontbinding (SVD)
- Verschillende similariteitsmaten (Weeds 2003)
- Verschillende wegingsfactors (Curran & Moens 2002)

Genereren van profielen (*synsets*)

Afstandsmeting:

- Matrix met voor elke lemma, vector van contextfeatures
- Eventueel Singuliere waardenontbinding (SVD)
- Verschillende similariteitsmaten (Weeds 2003)
- Verschillende wegingsfactors (Curran & Moens 2002)

Clustering:

- Lemma's moeten op basis van afstandsmeting gegroepeerd worden tot sets van synoniemen
- (Niet-hiërarchische) clusteringtechnieken
- Verschillende cut-off points
- Keywords als seeds

Genereren van profielen (*synsets*)

Evaluatie:

- Zijn de gevonden sets wel degelijk synoniemen?
 - Weerspiegelen de afstanden op basis van distributionele similariteit ook semantische afstanden?

Genereren van profielen (*synsets*)

Evaluatie:

- Zijn de gevonden sets wel degelijk synoniemen?
 - Weerspiegelen de afstanden op basis van distributionele similariteit ook semantische afstanden?
- Verschillende toetsstenen:

Genereren van profielen (*synsets*)

Evaluatie:

- Zijn de gevonden sets wel degelijk synoniemen?
 - Weerspiegelen de afstanden op basis van distributionele similariteit ook semantische afstanden?
- Verschillende toetsstenen:
 - Oorspronkelijke dataset met voetbal- en kledingstermen

Genereren van profielen (*synsets*)

Evaluatie:

- Zijn de gevonden sets wel degelijk synoniemen?
 - Weerspiegelen de afstanden op basis van distributionele similariteit ook semantische afstanden?
- Verschillende toetsstenen:
 - Oorspronkelijke dataset met voetbal- en kledingstermen
 - Lexicale ontologie met semantische relaties -> Eurowordnet
 - Correlatie met afstanden in ontologie
 - Dekking, kwaliteit? (Evaluatie van eurowordnet, zie Yves)

Genereren van profielen (*synsets*)

Evaluatie:

- Zijn de gevonden sets wel degelijk synoniemen?
 - Weerspiegelen de afstanden op basis van distributionele similariteit ook semantische afstanden?
- Verschillende toetsstenen:
 - Oorspronkelijke dataset met voetbal- en kledingstermen
 - Lexicale ontologie met semantische relaties -> Eurowordnet
 - Correlatie met afstanden in ontologie
 - Dekking, kwaliteit? (Evaluatie van eurowordnet, zie Yves)
 - Similariteitsoordelen van proefpersonen
 - data van natural-kind-concepten

Genereren van profielen (*synsets*)

Evaluatie:

- Zijn de gevonden sets wel degelijk synoniemen?
 - Weerspiegelen de afstanden op basis van distributionele similariteit ook semantische afstanden?
- Verschillende toetsstenen:
 - Oorspronkelijke dataset met voetbal- en kledingstermen
 - Lexicale ontologie met semantische relaties -> Eurowordnet
 - Correlatie met afstanden in ontologie
 - Dekking, kwaliteit? (Evaluatie van eurowordnet, zie Yves)
 - Similariteitsoordelen van proefpersonen
 - data van natural-kind-concepten
 - Selectie van significante contextfeatures voor a-priori set van synoniemen

Overzicht

1. Achtergrond:
 - Profielgebaseerde meting van lexicale variatie
2. sem-matrix en zijn deelprojecten
3. Genereren van profielen (*synsets*)
 - Data en dataverrijking
 - Contextmodel
 - Afstandsmeting en clustering
 - Evaluatie
4. Eerste resultaten
 - Betekenisverwante woorden vinden aan de hand van collocaten en dependentierelaties → Yves Peirsman

Overzicht

1. Achtergrond:
 - Profielgebaseerde meting van lexicale variatie
2. sem-matrix en zijn deelprojecten
3. Genereren van profielen (*synsets*)
 - Data en dataverrijking
 - Contextmodel
 - Afstandsmeting en clustering
 - Evaluatie
4. Eerste resultaten
 - Betekenisverwante woorden vinden aan de hand van collocaten en dependentierelaties → Yves Peirsman over hemden, corners en liefde



voor meer informatie:

kris.heylen@arts.kuleuven.be

yves.peirsman@arts.kuleuven.be

<http://www.ling.arts.kuleuven.be/qlvl/>

