Lexical convergence in the Dutch lexicon

Jocelyne Daems Kris Heylen Dirk Geeraerts

University of Leuven
RU Quantitative Lexicology and Variational Linguistics

Overview

1. Introduction

Variation in Dutch

Lectal variation

- register
- region
- gender
- generation
- profession and education

Linguistic variation

- vrachtwagen of vrachtauto (of camion of truck)
- dat ken/kan/kun jij wel zeggen
- zeven of zeuven

1. Introduction

Variation in Dutch

Lectal variation

- register
- region
- gender
- generation
- profession and education

Linguistic variation

- vrachtwagen of vrachtauto
- dat ken/kan jij wel zeggen
- zeven of zeuven

A specific context can come with its own set of linguistic choices
$\rightarrow$ lect

1. Introduction

Variation in Dutch *in this study*

Lectal variation

- region and register

$\Rightarrow$ geographical and stylistic/stratificational question

Linguistic variation

- lexicon

$\Rightarrow$ onomasiological question

1. Introduction

Research aim

Measure lexical convergence between Dutch in Belgium and Dutch in the Netherlands
by means of the word choice for emotion-, IT- and traffic concepts.

$\Rightarrow$ Cf. Geeraerts, Grondelaers & Speelman 1999

2. Dutch

Dutch

Two national varieties (pluricentric language (Clyne 1992))

- Netherlandic Dutch
- Belgian Dutch

Stratification

- Poldernederlands ?
- Verkavelingsvlaams, soapvlaams, tussentaal
  "in-between language"

2. Dutch

Standardisation

Different development of the standardisation process

- the Netherlands: indepdent development of a/the standard variant of Dutch
- Belgium: after a long delay, promotion of convergence with the (established) Netherlandic Dutch norm

Two opposed views in Belgium

- Integrationism: Netherlandic Dutch as the norm
- Particularism: room for a Belgian Dutch standard variant

3. Method

Geeraerts, Grondelaers & Speelman 1999

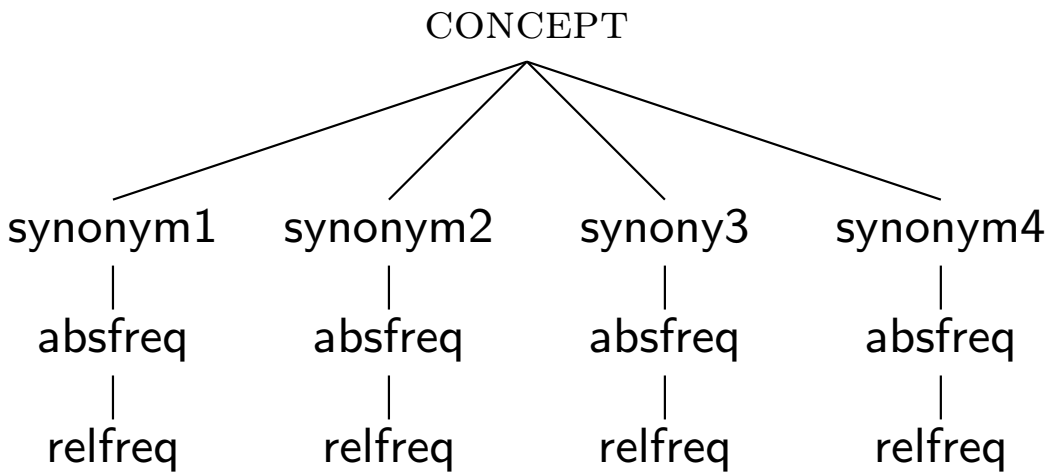
- uniformity measure for onomasiological variation

⇒ to what extent do people for a given concept choose the same words?

⇒ how often do Belgians and Dutchmen express VRACHTWAGEN ("truck") in the same way?

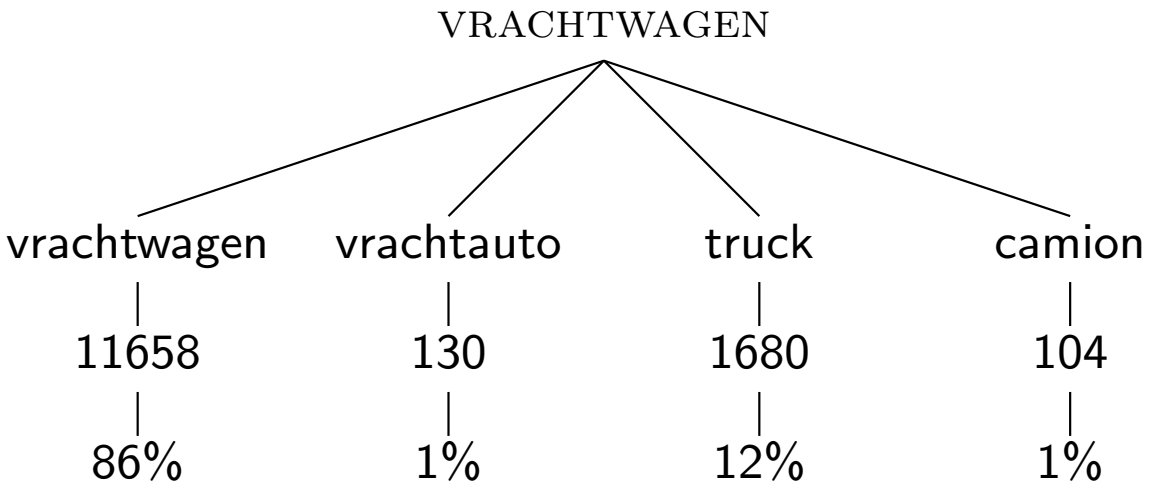
3. Method - Uniformity measure

Onomasiological profile



3. Method - Uniformity measure

Onomasiological profile: concrete



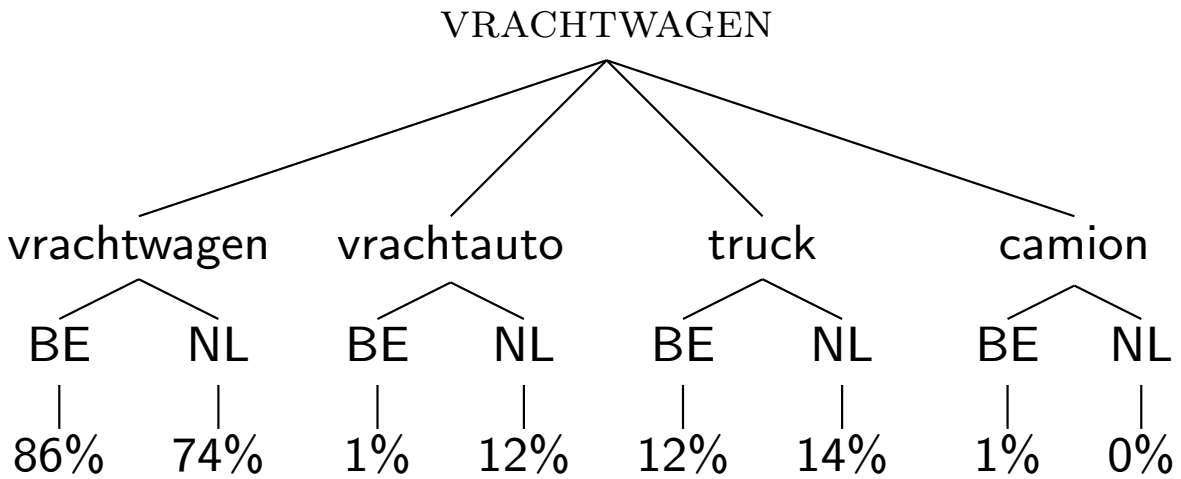
3. Method - Uniformity measure

Measure lexical convergence: five steps

1. Per concept build the onomasiological profile
2. Per variety build the onomasiological profile
3. Measure uniformity
4. Aggregate over all concepts
5. (Weigh the uniformity measure by means of the concept frequency)

3. Method - Uniformity measure

Measure lexical convergence: concrete



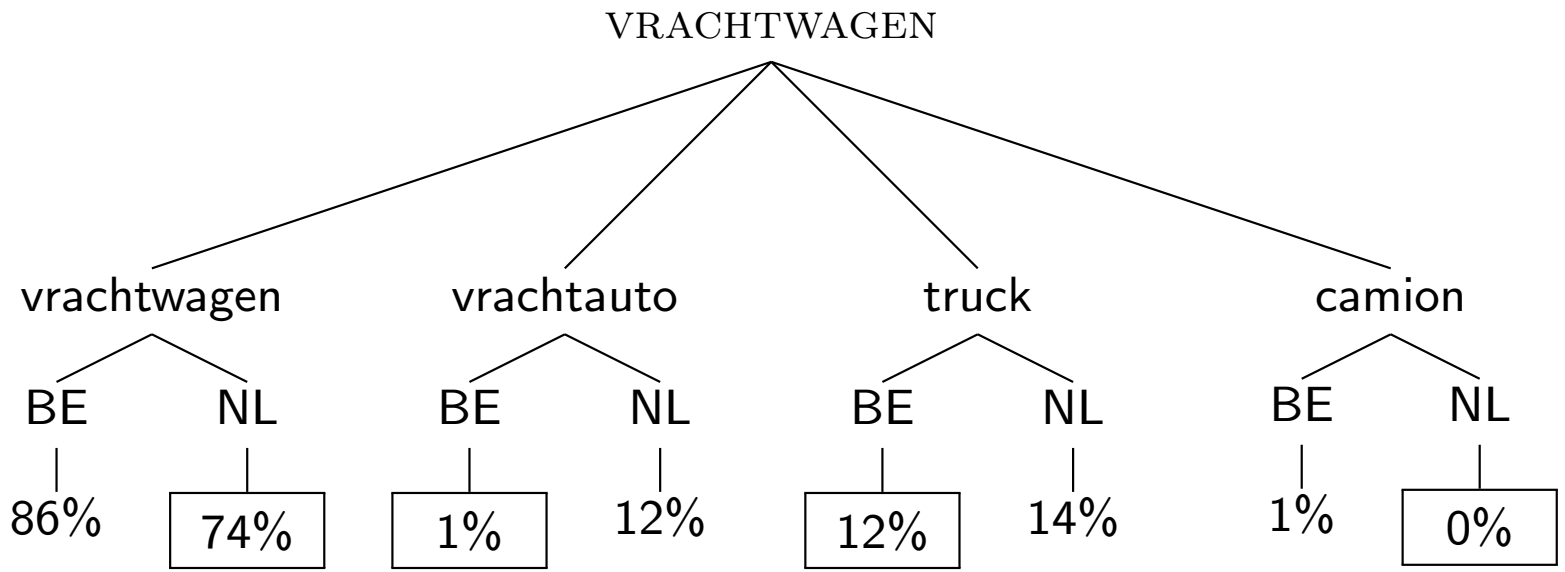
3. Method - Uniformity measure

Measure lexical convergence: five steps

1. Per concept build the onomasiological profile
2. Per variety build the onomasiological profile
3. Measure uniformity
   - $\Rightarrow$ = overlap in lexicalisation preferences
     = sum of the smallest relative value for each term
4. Aggregate over all concepts
5. (Weigh the uniformity measure by means of the concept frequency)

3. Method - Uniformity measure

Measure lexical convergence: concrete



$\Rightarrow 74\% + 1\% + 12\% + 0\% = 87\% \Leftarrow$ measure of uniformity

3. Method - Uniformity measure

Measure lexical convergence: five steps

1. Per concept build the onomasiological profile
2. Per variety build the onomasiological profile
3. Measure uniformity
4. Aggregate over all concepts
   $\Rightarrow$ = repeat for each concept in the semantic field
   + compute the mean
5. (Weigh the uniformity measure by means of the concept frequency)

3. Method - Uniformity measure

Measure lexical convergence: concrete

	U
VRACHTWAGEN	87%
METRO	97%
TREIN	100%
STOEP	61%
BESTELWAGEN	45%
mean	78%

3. Method - Uniformity measure

Measure lexical convergence: five steps

1. Per concept build the onomasiological profile
2. Per variety build the onomasiological profile
3. Measure uniformity
4. Aggregate over all concepts
5. Weigh the uniformity measure by means of the concept frequency

   $\Rightarrow$ = less frequent concepts weigh less

3. Method - Uniformity measure

Measure lexical convergence: concrete

	U	W		U'
VRACHTWAGEN	87%	32%		28%
METRO	97%	8%		7%
TREIN	99%	46%		45%
STOEP	61%	10%		6%
BESTELWAGEN	45%	4%		2%
mean	78%		sum	89%

3. Method - Data

Data gathering

Four steps

- Select profiles
- Lemmatised corpora
- Script frequencies and concordances
- Manual disambiguation

3. Method - Data

Profiles traffic

convertible
cabrio
cabriolet

fiets
rijwiel
velo
tweewieler

rondpunt
rotonde

pechstrook
vluchstrook

ondergrondse
metro
metrotrein

voetgangersoversteekplaats
zebra zebrapad

3. Method - Data

Profiles IT

mobiel gsm
mobieltje
mobiele telefoon

desktop
bureaublad

laptop
draagbare computer
schootcomputer
notebook

klavier
keyboard
toetsenbord

wachtwoord
paswoord
password

3. Method - Data

Profiles emotions

3. Method - Data

A sample of annotated text by Alpino

['Na', 'prep', 'na', 3849, 0], ['een', 'det', 'een', 3849, 1], ['jaar', 'noun', 'jaar', 3849, 2], ['-', 'punct', '-', 3849, 3], ['ik', 'pron', 'ik', 3849, 4], ['deed', 'verb', 'doe', 3849, 5], ['alles', 'noun', 'alles', 3849, 6], [',', 'punct', ',', 3849, 7], ['van', 'prep', 'van', 3849, 8], ['was', 'noun', 'was', 3849, 9], ['ophangen', 'verb', 'hang_op', 3849, 10], ['tot', 'prep', 'tot', 3849, 11], ['de', 'det', 'de', 3849, 12], ['beenhouwerij', 'noun', 'beenhouwerij', 3849, 13], ['kuisen', 'verb', 'kuis', 3849, 14], ['-', 'punct', '-', 3849, 15], ['vroeg', 'verb', 'vraag', 3849, 16], ['ik', 'pron', 'ik', 3849, 17], ['of', 'comp', 'of', 3849, 18], ['ik', 'pron', 'ik', 3849, 19], ['meer', 'det', 'meer', 3849, 20], ['kon', 'verb', 'kan', 3849, 21], ['verdienen', 'verb', 'verdien', 3849, 22], ['.', 'punct', '.', 3849, 23], ['De', 'det', 'de', 3850, 0], ['slager', 'noun', 'slager', 3850, 1], ['wees', 'verb', 'wijs_af', 3850, 2], ['dat', 'det', 'dat', 3850, 3], ['verzoek', 'noun', 'verzoek', 3850, 4], ['af', 'part', 'af', 3850, 5], ['.', 'punct', '.', 3850, 6], ['Ik', 'pron', 'ik', 3851, 1], ['pakte', 'verb', 'pak', 3851, 2], ["m'n", 'det', 'mijn', 3851, 3], ['velo', 'noun', 'velo', 3851, 4], ['en', 'vg', 'en', 3851, 5], ['ik', 'pron', 'ik', 3851, 6], ['was', 'verb', 'ben', 3851, 7], ['weg', 'adv', 'weg', 3851, 8], ['.', 'punct', '.', 3851, 9]

⇒ Na een jaar – ik deed alles, van was ophangen tot de beenhouwerij kuisen – vroeg ik of ik meer kon verdienen. De slager wees dat verzoek af. Ik pakte m'n velo en ik was weg.

3. Method - Data

Manual desambiguation

Stevaert vindt de sluiting van de kleine ring rondom Antwerpen een topprioriteit. ⇒ relevant meaning

Al draagt een aap een gouden ring, het is en blijft een lelijk ding. ⇒ wrong meaning: jewellery

Wat is de kick ? Als ik in de ring oog in oog sta met mijn tegenstander. ⇒ wrong meaning: boxing ring

3. Method - Data

Profiles

80 concepts

2 parts of speech

3 lexical fields

- traffic: VRACHTWAGEN ("truck")
- IT: GSM ("cell phone")
- emotions: VERLEGEN ("shy"), OPRECHTHEID ("sincerity")

Corpora

	newspaper	Usenet
Belgium	373M	18M
the Netherlands	161M	24M

4. Results

Hypotheses

1. Diachronic convergence
   - = word choice in Belgium and the Netherlands becomes more uniform
2. Synchronic stratification
   - = word choice in the Netherlands is more uniform than in Belgium
   - = word choice in newspapers is more uniform than in Usenet
3. Role of the semantic field
4. Role of the concept frequency

4. Results

Hypothesis 1: Diachronic convergence

Study : Geeraerts, Grondelaers & Speelman 1999
Data : clothing terms in fashion magazines

$U'_{(B50,N50)}$	70%
$U'_{(B70,N70)}$	75%
$U'_{(B90,N90)}$	82%

Study : De Cnodder 2013
Data : clothing terms in fashion magazines

$U'_{(B12,N12)}$	78%

4. Results

Hypothesis 1: Diachronic convergence

Study : Geeraerts, Grondelaers & Speelman 1999

Data : football terms in magazines in 1990

$U'_{(B50,N50)}$	66%
$U'_{(B70,N70)}$	72%
$U'_{(B90,N90)}$	77%

4. Results

Hypothesis 1: Diachronic convergence

Study : Grondelaers, Van Aken, Speelman & Geeraerts 2001
Data : clothing terms in newspapers in 1998

$U'_{(B58,N58)}$	71%
$U'_{(B78,N78)}$	64%
$U'_{(B98,N98)}$	82%

Study : Grondelaers, Van Aken, Speelman & Geeraerts 2001
Data : prepositions in newspapers in 1998

$U'_{(B58,N58)}$	59%
$U'_{(B78,N78)}$	62%
$U'_{(B98,N98)}$	67%

4. Results

Hypothesis 1: Diachronic convergence

- Convergence for the supraregional/formal language use
  50 – 70 – 90 / 58 – 78 – 98
- Shift in Belgian Dutch, which converges towards Netherlandic Dutch

! Situation in 2012: stagnation or divergence?
  - Informalisation of Netherlandic Dutch?

! Balanced composition of the corpus

! Content words vs. function words

4. Results

Hypothesis 2: Synchronic stratification

Study : Geeraerts, Grondelaers & Speelman 1999

Data : clothing terms in fashion magazines and shop windows in 1990

U'$_{(B_{mag},B_{etal})}$	50%
U'$_{(N_{mag},N_{etal})}$	69%

Study : De Cnodder 2013

Data : clothing terms in fashion magazines and shop windows in 2012

U'$_{(B_{mag},B_{etal})}$	63%
U'$_{(N_{mag},N_{etal})}$	79%

4. Results

Hypothesis 2: Synchronic stratification

Study : Grondelaers, Van Aken, Speelman & Geeraerts 2001
Data : clothing terms in newspapers, Usenet and IRC in 1998

$U'_{(B_{krant}, B_{Usenet})}$	90%
$U'_{(B_{krant}, B_{IRC})}$	83%
$U'_{(N_{krant}, N_{Usenet})}$	92%
$U'_{(N_{krant}, N_{IRC})}$	90%

Study : Grondelaers, Van Aken, Speelman & Geeraerts 2001
Data : prepositions in newspapers, Usenet and IRC in 1998

$U'_{(B_{krant}, B_{Usenet})}$	82%
$U'_{(B_{krant}, B_{IRC})}$	75%
$U'_{(N_{krant}, N_{Usenet})}$	97%
$U'_{(N_{krant}, N_{IRC})}$	94%

4. Results

Hypothesis 2: Synchronic stratification

Study : Daems, Heylen & Geeraerts (In prep.)
Data : traffic terms in newspapers and Usenet in 1999-2005

$U_{(B_{krant}, B_{Usenet})}$	87%
$U_{(N_{krant}, N_{Usenet})}$	86%

Study : Daems, Heylen & Geeraerts (In prep.)
Data : IT terms in newspapers and Usenet in 1999-2005

$U_{(B_{krant}, B_{Usenet})}$	72%	$\Rightarrow$	*mailinglist*, *provider*, *database*
$U_{(N_{krant}, N_{Usenet})}$	70%	$\Rightarrow$	*mailinglist*, *provider*, *password*

4. Results

Hypothesis 2: Synchronic stratification

- Uniformity is lower in Belgian Dutch
- In Belgian Dutch there is a continuum from regional/informal (IRC) to supraregional/formal (newspapers) language use in which Usenet takes an intermediary position

! Importance semantic field

4. Results

Hypothesis 2: Synchronic stratification (bis)

Studies : Geeraerts, Grondelaers & Speelman 1999

Data : clothing terms in fashion magazines in 1990 and 2012, newspapers, Usenet and IRC in 1998

$U'_{(B_{mag,90},N_{mag,90})}$	82%
$U'_{(B_{mag,12},N_{mag,12})}$	78%
$U'_{(B_{krant},N_{krant})}$	82%
$U'_{(B_{Usenet},N_{Usenet})}$	73%
$U'_{(B_{IRC},N_{IRC})}$	75%

4. Results

Hypothesis 2: Synchronic stratification (bis)

Study : Daems, Heylen & Geeraerts (In prep.)
Data : traffic terms in newspapers and Usenet in 1999-2005

$U'_{(B_{krant}, N_{krant})}$	78%		
$U'_{(B_{Usenet}, N_{Usenet})}$	72%	$\Rightarrow$	BUS, TREIN, VRACHTWAGEN

Study : Daems, Heylen & Geeraerts (In prep.)
Data : IT terms in newspapers and Usenet in 1999-2005

$U'_{(B_{krant}, N_{krant})}$	74%		
$U'_{(B_{Usenet}, N_{Usenet})}$	91%	$\Rightarrow$	WEBSITE, LINK, PROVIDER

4. Results

Hypothesis 2: Synchronic stratification (bis)

- On a supraregional level the standard variant is more uniform than the informal variant(s)

! Importance of the semantic field

4. Results

Hypothesis 3: Role of the semantic field

Studies : Geeraerts, Grondelaers & Speelman 1999, Grondelaers, Van Aken, Speelman & Geeraerts 2001

Data : football terms in magazines in 1990, prepositions and clothing terms in newspapers in 1998

$U'_{(B_{kleding}, N_{kleding})}$	82%
$U'_{(voorzetsel, N_{voorzetsel})}$	67%
$U'_{(B_{voetbal}, N_{voetbal})}$	77%

4. Results

Hypothesis 3: Role of the semantic field

Study : Daems, Heylen & Geeraerts (In prep.)

Data : traffic, IT and emotion terms in newspapers in 1999-2005

$U'_{(B_{verkeer}, N_{verkeer})}$	78%
$U'_{(B_{ICT}, N_{ICT})}$	74%
$U'_{(B_{emotie,noun}, N_{emotie,noun})}$	81%
$U'_{(B_{emotie,adj}, N_{emotie,adj})}$	79%

4. Results

Hypothesis 3: Role of the semantic field

- Age of the field, degree of supraregional contact, register

! Statistic significance: requires enough data

4. Results

Hypothesis 4: Role of the concept frequency

Study : Daems, Heylen & Geeraerts (In prep.)

Data : traffic, IT and emotion terms in newspapers in 1999-2005

	non-weighted (U)	weighted (U')
$(B_{verkeer}, N_{verkeer})$	70%	78%
(B_{ICT}, N_{ICT})	79%	74%
$(B_{emotie,noun}, N_{emotie,noun})$	83%	81%
$(B_{emotie,adj}, N_{emotie,adj})$	78%	79%

4. Results

Hypothesis 4: Role of the concept frequency

Study : Daems, Heylen & Geeraerts (In prep.)
Data : traffic, IT and emotion terms in newspapers in 1999-2005

	non-weighted (U)	weighted (U')
($B_{verkeer}$,$N_{verkeer}$)	70%	78%
(B_{ICT},N_{ICT})	79%	74%

traffic 70% $\rightarrow$ 78% $<$ BUS, FIETS, TREIN, VRACHTWAGEN
IT 79% $\rightarrow$ 74% $<$ MOBIELE TELEFOON, E(-)MAIL

4. Results

Hypothesis 4: Role of the concept frequency

- Know your data qualitatively

5. Conclusion

Confirmation

- Diachronic convergence
- Synchronic stratification

Attention

- ! Importance of the semantic field
- ! Importance of the concept frequency

5. Conclusion

Future

Semasiological perspective

- Correlation between onomasiological and semasiological perspective

Automation: techniques from distributional semantics

- Vector Space Models (VSMs)
  - Type-based VSMs
  - Token-based VSMs

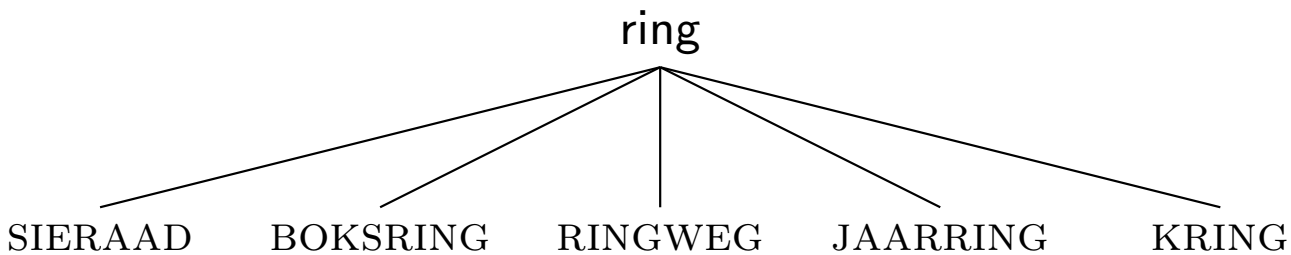
5. Conclusion

Future - Semasiological perspective

Onomasiology: given a concept - how is it named?

⇒ Semasiology: given a word - what does it mean?

Bv. AUTOSNELWEG vs. RING



5. Conclusion

Future - Semasiological perspective

Case study traffic:

> Onomasiologically a higher uniformity between Belgian Dutch and Netherlandic Dutch than semasiologically,
>
> in other words, we rather name things the same than that words have the same meaning.

5. Conclusion

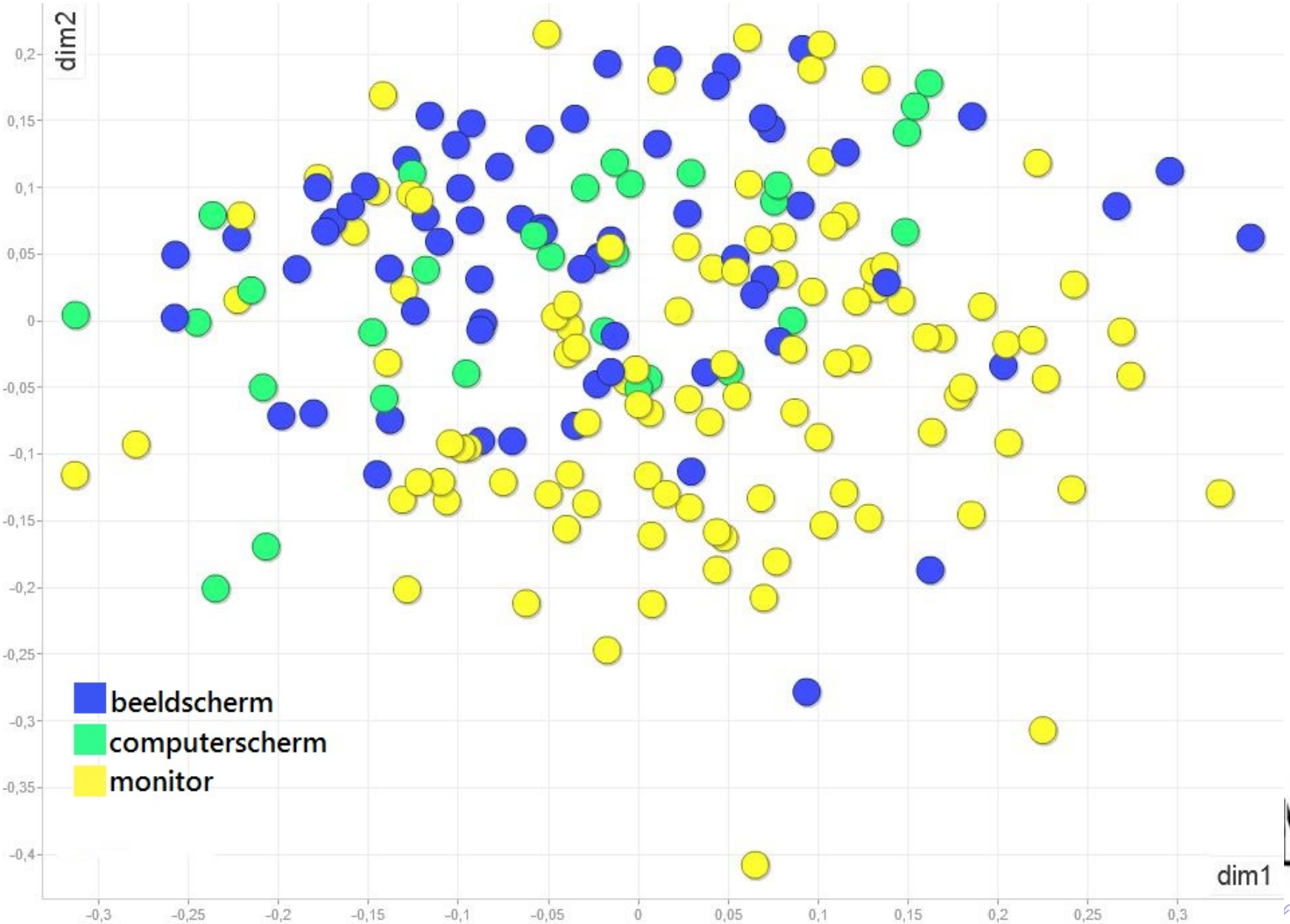
Future - Distributional semantics - Vector Space Models

"You shall know a word by the company it keeps" (Firth)

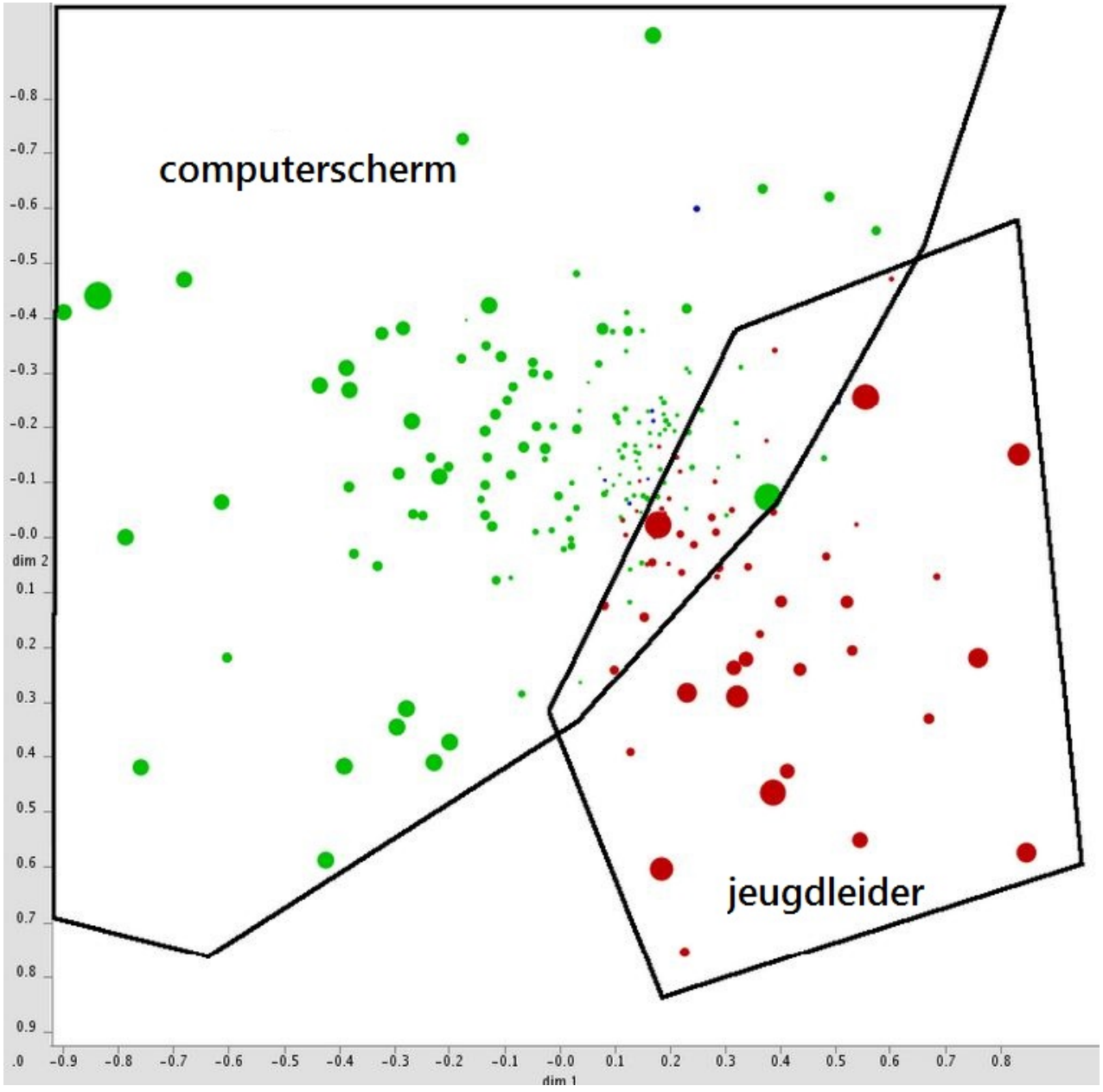
Type-based ˜ onomasiology: detect synonyms

Token-based ˜ semasiology: detect senses

5. Conclusion



5. Conclusion



Thank you! Questions?

Further information?

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

jocelyne.daems@arts.kuleuven.be

References 1/2

- Bouma, Gosse, Gertjan van Noord & Rob Malouf. 2001 "Alpino: wide-coverage computational analysis of Dutch." In: Walter Daelemans, et al. (eds). *Computational Linguistics in the Netherlands 2000*. Amsterdam: Rodopi, 45-59.
- Clyne, Michael. 1992. *Pluricentric languages: differing norms in different nations.* Berlin/New York: Mouton de Gruyter.
- De Cnodder, Tine. 2013. *Convergentie en divergentie in de Nederlandse woordenschat anno 2012.* Unpublished MA thesis. KU Leuven.
- Geeraerts, Dirk, Stefan Grondelaers & Dirk Speelman. 1999. *Convergentie en divergentie in de Nederlandse woordenschat: een onderzoek naar kleding- en voetbaltermen.* Amsterdam: P.J. Meertens-Instituut.
- Grondelaers, Stefan, Dirk Geeraerts, Dirk Speelman and José Tummers. 2001. Lexical standardisation in internet conversations. Comparing Belgium and The Netherlands. In: Fontana, McNally, Turell and Enric Vallduví (eds.), *Proceedings of the First International Conference on Language Variation in Europe*, 90–100. Barcelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada, Unitat de Investigació de Variació Lingüística.

References 2/2

- Grondelaers, Stefan, Hilde Van Aken, Dirk Speelman & Dirk Geeraerts. 2001. Inhoudswoorden en prepositie als standaardiseringsindicatoren. De diachrone en synchronic status van het Belgische Nederlands. *Nederlandse Taalkunde* 6: 179–202.
- Heylen, Kris, Dirk Speelman & Dirk Geeraerts. 2012. "Looking at word meaning. An interactive visualization of Semantic Vector Spaces for Dutch synsets." In Miriam Butt, Sheelagh Carpendale, Gerald Penn, Jelena Prokic & Michael Cysouw (eds.), *Proceedings of the EACL-2012 joint workshop of LINGVIS & UNCLH: Visualization of Language Patters and Uncovering Language History from Multilingual Resources*, 1624. Stroudsburg: Association for Computational Linguistics.
- Speelman, Dirk, Stefan Grondelaers & Dirk Geeraerts. 2003. "Profile-based linguistic uniformity as a generic method for comparing language varieties." *Computers and the Humanities* 37 (3): 317-337.
- Wielfaert, Thomas, Kris Heylen & Dirk Speelman. 2013. "Visualisations interactives des espaces vectoriels sémantiques pour l'analyse lexicologique." *Actes de SemDis 2013 : Enjeux actuels de la sémantique distributionnelle*:154-166.