Explaining lexical diversity in dialect data: The influence of concept features

Karlien Franco
– PhD researcher at University of Leuven (QLVL)
– supervisors:

Dirk Geeraerts    Roeland van Hout    Dirk Speelman

Concept features and lexical diversity

A dialectological case study on the relationship between meaning and variation

Karlien Franco

Diversity in Language, Culture & Cognition Colloquium series, 09.11.2017
lexical diversity across languages

= number of different words to refer to a particular concept

• variation:
  e.g. words for SNOW (cf. Boas 1911, Regier et al. 2016):
    Eskimo languages: *aput, qana, sirpoq*
    English: *snow*
  words for colors
    (e.g. Berlin & Kay 1969, Kay & Regier 2006)
  words for the senses
    (e.g. Majid & Burenhult 2014)
  ...

• stability < biological, cognitive, ... characteristics

• variation < differences in *conceptualization* (& lexicalization)
socio-cultural environment
categorization depends “upon the chief interests of a people; and where it is necessary to distinguish a certain phenomenon in many aspects, which in the life of the people play each an entirely independent role, many independent words may develop, while in other cases modifications of a single term may suffice”
(Boas 1911: 26)

“language “entrenching” cognitive differences induced by cultural embodiment and cultural practice.”
(Sinha & Jensen de Lopez 2000: 37)

“The languages spoken today showcase the diverging sociocultural, environmental and linguistic histories each language has undergone.”
(Malt & Majid 2013: 592)
variation in lexical diversity within one language

e.g. **SOBER**:

sober, abstinent

**DRUNK**:

drunk, blitzed, intoxicated,
hammered, pissed, canned,
I’m not as think as you drunk I am, ...

e.g. Brabantic dialects:

**SLUIS** ‘lock (shipping)’:

sluis, sas, ... (5 variants)

**IEMAND WEERSTAAN** ‘to resist (someone)’:

bolwerken, volhouden, niet toegeven, ... (51 variants)
distribution of lexical items

- not only **number of different variants** but also how these variants are distributed in different **lects**
  - one word for every situation?
  - different words depending on speaker or situation? (Geeraerts, Grondelaers & Speelman 1999)

- **lectal variation**
  - **sociolectal** e.g. *blitzed* vs. *intoxicated*
  - **regiolectal** e.g. *pissed* (UK) vs. *canned* (US)
  - **dialectal**: spatial patterns of variation
  - ...
taking into account (dia)lectal variation

SLUIS
‘lock (shipping)’

IEMAND WEERSTAAN
‘to resist (someone)’
(N > 5)
meaning influences dialectal diversity

- background
- data
- two case-studies
- conclusions
BACKGROUND
traditional explanations of variation in lexical diversity

- **taboo concepts** show more variation (Allan & Burridge 1988, 2006):
  → prone to rapid language change because euphemistic meaning is quickly lost

- **core vocabulary** is not prone to change or borrowing
variation in lexical diversity in dialectology

• lack of lexical diversity ~ dialect levelling
• dialect levelling < geography and social or political factors
e.g. mobility, population size, types of language learning,
presence of state border, differences in cultural practices, ... 

• Pickl (2013) and Szelid & Geeraerts (2008): social and semantic features cause variation in the homogeneous spatial diffusion of variants

• generally small-scale; not comprehensively researched
Cognitive Sociolinguistics

- convergence between Cognitive Linguistics and sociolinguistics
- social nature of language
- maximalist perspective on meaning
  → prototype-theoretical organization of the lexicon
  - no necessary and sufficient conditions
  - differences in degree of membership and typicality

Kristiansen & Dirven (2008)
Geeraerts, Kristiansen & Peirsman (2010)
Pütz, Robinson & Reif (2014)
two crucial distinctions

- semasiology vs. onomasiology
- formal vs. conceptual variation

Geeraerts, Grondelaers & Bakema (1994)
semasiology vs. onomasiology

**semasiology** ~ study of meaning
- word → meanings that word can take
  e.g. *blitzed* → DRUNK, WORN OUT

**onomasiology** ~ study of naming
- concept → words for the concept
  e.g. *SNOW* → *snow*
- onomasiological profile: all words for a concept + relative frequency (Geeraerts, Grondelaers & Speelman 1999)
  e.g. *SOBER* (61 mio.): *sober* 88%
    abstinrent 12%
formal onomasiological variation

“jeans”

“blue jeans”

“denims”
conceptual onomasiological variation

“jeans”, “blue jeans”, “denims”

“pants”, “trousers”
focus on formal variation

“jeans”, “blue jeans”, “denims”

→ lexical diversity as the number of synonymous expressions & their lectal distribution
lexical variation in Cognitive Sociolinguistics

interacts with prototype-theoretical organization of the lexicon

- e.g. typicality differences: OVERHEMD ‘dress shirt’

→ “overhemand”

is more typical than
pilot studies on lexical diversity

• first to directly test correlation between semantic features and lexical diversity
  - negative affect
  - onomasiological vagueness
  - lack of onomasiological salience

• limited to one dialect area & one semantic field

→ are semantic concept features also influential in other semantic fields and other dialect areas?

Geeraerts & Speelman (2010), Speelman & Geeraerts (2008)
Woordenboek van de Brabantse dialecten
‘Dictionary of the Brabantic dialects’

Woordenboek van de Limburgse dialecten
‘Dictionary of the Limburgish dialects’
WBD & WLD

- onomasiological dialect dictionaries
- digitized databases
- one or more semantic fields per case-study
- one volume = one semantic field
PART 3: General vocabulary

1: Man as an individual
- The human body
- Physical activity and health
- Clothing and grooming
- Personality and feelings

2: Domestic life
- The house
- Family and sexuality
- Food and drink

3: Community life
- Society, school and education
- Celebration and entertainment
- Church and religion

4: The world versus man
- Fauna: birds
- Fauna: other animals
- Flora
- The physical and abstract world
data included in the analyses

- questionnaire data
- systematically elicited data

- data as it is available
  - no manual changes
  - advantage: large data set
  - disadvantage: little control over data
    e.g. different number of observations per concept, no specific information about background of respondents
  → aggregation
from questionnaire...

KEUKEN-EN KOOKGEREI

2.
1. Hoe noemt u in uw dialect:
   Het metalen voorwerp met heengel en tuit dat dient om water in te koken?
   a. het moderne voorwerp (zie ill.) ........................................

2. b. het oudere voorwerp, dikwijls met ronde bodem, dat boven een haardvuur werd gehangen of in de opening van een kachel werd gezet (zie ill.) ........................................

3. De pot met deksel, tuit en oor waarin koffie wordt gezet?
   ........................................

4. De zak waardoor het water over de gemalen koffie wordt gegoten?
   ........................................

5. De buikige pot met tuit en oor waarin enkel gezet?
   ........................................

6. Wanneer werd er thee gedronken? Als dagelijk als geneeskrachtige drank?

WOORDENBOEK VAN DE VLAAMSE DIALEKten : bijlage bij WVD 43 : huisraad
(Gelieve de illustraties niet terug te sturen)
Het nummer van de illustratie komt overeen met het nummer van de vraag.

Diversity in Language, Culture & Cognition Colloquium series, 09.11.2017
<table>
<thead>
<tr>
<th>concept</th>
<th>variant (dutchified form)</th>
<th>question</th>
<th>location</th>
</tr>
</thead>
<tbody>
<tr>
<td>damesmantel</td>
<td>caban (fr.)</td>
<td>damesmantel, inventarisatie uitdrukkingen</td>
<td>Tervuren</td>
</tr>
<tr>
<td>‘coat for women’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>overjas</td>
<td>frak</td>
<td>een jas die men over het colbert heen draagt</td>
<td>Leopoldsburg</td>
</tr>
<tr>
<td>‘overcoat’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vrolijk</td>
<td>spass (du.) haan</td>
<td>een opgeruimde, lichte, blijde stemming [...]</td>
<td>Simpelveld</td>
</tr>
<tr>
<td>‘cheerful’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vrolijk</td>
<td>opgewekt</td>
<td>een opgeruimde, lichte, blijde stemming [...]</td>
<td>Venlo</td>
</tr>
<tr>
<td>‘cheerful’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

... to dataset ...

Diversity in Language, Culture & Cognition Colloquium series, 09.11.2017
... to measurements at the level of the concept

<table>
<thead>
<tr>
<th>concept</th>
<th>lexical diversity</th>
<th>predictor 1: affect sensitivity</th>
<th>predictor 2: vagueness</th>
</tr>
</thead>
<tbody>
<tr>
<td>achterdochtig</td>
<td>5</td>
<td>sensitive</td>
<td>2.275</td>
</tr>
<tr>
<td>‘suspicious’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>achterhoofd</td>
<td>21</td>
<td>neutral</td>
<td>4.977</td>
</tr>
<tr>
<td>‘back of the head’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>speelplaats</td>
<td>3</td>
<td>neutral</td>
<td>2.341</td>
</tr>
<tr>
<td>‘playground’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>speels</td>
<td>9</td>
<td>sensitive</td>
<td>3.561</td>
</tr>
<tr>
<td>‘light-hearted’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

Diversity in Language, Culture & Cognition Colloquium series, 09.11.2017
NB: phonological variation
1. SCALING UP

Do the semantic features have a significant and stable effect in other semantic fields and dialect areas?

Franco, Geeraerts, Speelman & Van Hout (in prep.)
<table>
<thead>
<tr>
<th>pilot study</th>
<th>this replication study</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 semantic field</td>
<td>6 semantic fields</td>
</tr>
<tr>
<td>the human body</td>
<td>concrete – abstract</td>
</tr>
<tr>
<td></td>
<td>individual – local community – society</td>
</tr>
<tr>
<td>1 dialect area</td>
<td>2 dialect areas</td>
</tr>
<tr>
<td>Limburgish dialects</td>
<td>Limburgish &amp; Brabantic dialects</td>
</tr>
<tr>
<td></td>
<td>$N_{\text{concepts}} = 3136$</td>
</tr>
<tr>
<td></td>
<td>$N_{\text{locations}} = 660$</td>
</tr>
<tr>
<td></td>
<td>$N_{\text{source}} = 532,627$</td>
</tr>
</tbody>
</table>

identical semantic features & response variable
response variable: lexical diversity

composite variable:
  log(number of types * geographical fragmentation)

- number of types per concept
- geographical fragmentation per concept
  - lectal variation
  - quantification of spatial nature of dialectal data
  - more or less heterogeneous
Diversity in Language, Culture & Cognition Colloquium series, 09.11.2017

less heterogeneous

more heterogeneous

SLUIS ‘lock (shipping)’

IEMAND WEERSTAAN ‘to resist (someone)’

(N > 5)
explanatory variables: semantic features

- onomasiological vagueness
- lack of onomasiological salience
- proneness to affect

cf. Geeraerts et al. (1994): prototype-theoretical organization of the lexicon (vagueness & salience) interacts with the structure of lexical variation
vagueness

- no necessary and sufficient conditions
- differences in degree of membership
  - e.g. ‘potato’ for category ‘vegetables’
  - ‘necklace’ for category ‘clothing’
- also on the onomasiological level
  - → fuzziness between concepts belonging to the same semantic field

Rosch (1978, 1987)
Rosch & Mervis (1975)
Geeraerts, Grondelaers & Bakema (1994)
onomasiological vagueness

→ fuzziness at the edges
onomasiological vagueness

“cacti”
onomasiological vagueness

type of cactus

other type of succulent (Euphorbia)
hypothesis & operationalization

- onomasiologically vague concept ~ more lexical diversity
  < demarcation differences (Pickl 2013)
  cf. cacti: laymen vs. botanists

- operationalization: lexical non-uniqueness
  = number of lexical items that are also used for other concepts
  e.g. TO RAIN HEAVILY
    “to rain” → how often also used for other concepts?
    “to pour” (e.g. TO STORM)
onomasiological salience

how familiar is the concept for a dialect speaker?

• extension of basic-level theory

• the basic level, i.e. the generic taxon, constitutes the core of a folk-biological classification

“[a]t this rank, both plants and animals appear perceptually most distinct to the human classifier, and these differences in morphology and behaviour virtually ‘cry out to be named’”
(Berlin 1978: 24)
“oak”

vs. “tree”
vs. “blackjack oak”
“robin”

vs. “bird”
vs. “European robin”
two problems with basic-level theory

1. assumes strict taxonomical organization

DAMESKLEDINGSTUK ‘piece of clothing for women’

Geeraerts et al (1994)
two problems with basic-level theory

1. assumes strict taxonomical organization

2. differences in typicality between items at the same level
e.g. OAK vs. GREVILLEA (native to Australia)
generalized onomasiological salience

- onomasiological salience as “conceptual ‘wiring in’”: “a well-entrenched concept is more firmly anchored in the language user’s knowledge of the language.” (Geeraerts et al. 1994: 145)

- allows for differences in typicality/familiarity between items that can (but need not) be situated at the same level

- pilot studies: lack of onomasiological salience ~ more lexical diversity

Geeraerts et al (1994)
operationalization

- proportion of multi-word expressions
- proportion of missing places
- proportion of hapax legomena
- prevalence (Keuleers et al. 2015)
  - word-level
  - missing data
  → prevalent vs. not-prevalent/missing
affect

• taboo concepts show more lexical diversity
e.g. DRUNK – SOBER

• more generally:
  language users have clear positive or negative associations
  with words denoting particular concepts
  cf. Osgood & Tannenbaum (1957)

• pilot studies:
  negative affect ~ more lexical diversity
affect: operationalization

• forced-choice task:
  – use best judgement in deciding whether a particular concept has a connotation
  – negative, positive, neutral, unknown
• variable used in analyses: proportion of non-neutral ratings
  – certainty of non-neutrality
  – e.g. LIJKWAGEN ‘hearse’: 4 negative, 1 neutral rating
  → proportion of non-neutral ratings = 0.8

• only 5 respondents, but consistent with large-scale affect measurements (Moors et al 2013)
• moderate to substantial inter-rater reliability:
  Light’s kappa = 0.675
testing systematicity of semantic features

- two additional variables in regression model:
  - dialect area: Limburgish / Brabantic
  - six semantic fields

- hypothesis: semantic features stable across semantic fields & across dialect areas
six semantic fields

high average degree of concreteness

- the human body
e.g. HEAD, KNEE, FOOT, CORPULENT
- the house
e.g. CUTLERY, TYPES OF POTS, CLEANING UP, WASHING
- celebration & entertainment
e.g. sports & (children’s) games, celebrations (e.g. CARNIVAL), the arts (e.g. SCULPTOR)

(mean concreteness: Brysbaert et al. 2014)
six semantic fields

low average degree of concreteness

• personality & feelings
  e.g. (temporary) feelings (e.g. ANGER), personality traits (e.g. TO BE SHY), behaviour (e.g. HASTY), memory & thinking (e.g. TO INFORM)

• family & sexuality
  e.g. baptism, marriage, death and burial

• society, school & education
  e.g. police, war and defence, communication, schooling, transportation

(mean concreteness: Brysbaert et al. 2014)
### Semantic Fields Along Two Dimensions

<table>
<thead>
<tr>
<th></th>
<th>Concrete</th>
<th>Abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Semantic Field</strong></td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Individual</td>
<td>the human body</td>
<td>personality &amp; feelings</td>
</tr>
<tr>
<td>Locally-Bound</td>
<td>the house</td>
<td>family &amp; sexuality (WLD only)</td>
</tr>
<tr>
<td>Societal</td>
<td>celebration &amp; entertainment</td>
<td>society, school &amp; education</td>
</tr>
</tbody>
</table>

<lay-out dictionary>
RESULTS
linear regression model

- $R^2 = 0.7311$
- semantic concept features all have significant and expected effect
- no interaction effects with dialect area
- interactions with semantic field: some features have larger effect in particular semantic fields
<table>
<thead>
<tr>
<th>model term</th>
<th>estimate</th>
<th>SE</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>2.586</td>
<td>0.072</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>dictionary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WBD</td>
<td>0.184</td>
<td>0.032</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td><strong>semantic field</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the house</td>
<td>0.344</td>
<td>0.082</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>celebration &amp; entertainment</td>
<td>0.059</td>
<td>0.079</td>
<td>NS</td>
</tr>
<tr>
<td>personality &amp; feelings</td>
<td>0.200</td>
<td>0.090</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>family &amp; sexuality</td>
<td>0.132</td>
<td>0.121</td>
<td>NS</td>
</tr>
<tr>
<td>society, school &amp; education</td>
<td>0.274</td>
<td>0.072</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td><strong>lack of salience</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>proportion of missing places</td>
<td>-1.055</td>
<td>0.104</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>proportion of MWE’s</td>
<td>0.583</td>
<td>0.076</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>proportion of hapaxes</td>
<td>13.318</td>
<td>0.552</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>prevalence binary (missing / not prevalent)</td>
<td>0.228</td>
<td>0.032</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td><strong>vagueness</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lexical non-uniqueness</td>
<td>0.032</td>
<td>0.003</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td><strong>affect</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>proportion of non-neutral ratings</td>
<td>0.280</td>
<td>0.042</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td><strong>interaction terms</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sem. field (the house) : proportion of hapaxes</td>
<td>1.483</td>
<td>0.792</td>
<td>&lt; 0.1</td>
</tr>
<tr>
<td>sem. field (celebration &amp; entertainment) : prop. of hapaxes</td>
<td>-3.220</td>
<td>0.638</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>sem. field (personality &amp; feelings): proportion of hapaxes</td>
<td>-1.867</td>
<td>0.626</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>sem. field (family &amp; sexuality) : proportion of hapaxes</td>
<td>0.736</td>
<td>1.205</td>
<td>NS</td>
</tr>
<tr>
<td>sem. field (society, school &amp; education) : prop. of hapaxes</td>
<td>-1.195</td>
<td>0.639</td>
<td>&lt; 0.1</td>
</tr>
<tr>
<td>sem. field (the house) : lexical non-uniqueness</td>
<td>-0.002</td>
<td>0.004</td>
<td>NS</td>
</tr>
<tr>
<td>sem. field (celebration &amp; entertainment) : lexical non-uniqueness</td>
<td>0.018</td>
<td>0.006</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>sem. field (personality &amp; feelings): lexical non-uniqueness</td>
<td>-0.012</td>
<td>0.003</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>sem. field (family &amp; sexuality) : lexical non-uniqueness</td>
<td>-0.007</td>
<td>0.010</td>
<td>NS</td>
</tr>
<tr>
<td>sem. field (society, school &amp; education) : lexical non-uniqu.</td>
<td>0.001</td>
<td>0.003</td>
<td>NS</td>
</tr>
<tr>
<td>proportion of hapaxes : lexical non-uniqueness</td>
<td>-0.053</td>
<td>0.007</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
onomasiological vagueness

STUNTELEN ‘to fumble’
vague (158)
N = 288

BLIND ‘blind’
not vague (0)
N = 327

Diversity in Language, Culture & Cognition Colloquium series, 09.11.2017
STUNTELEN ‘to fumble’ (N > 10)

- frotten (85)
- stuntelen (39)
- prutsen (36)
- haspelen (16)
- sukkelen (12)
- prullen (11)
onomasiological salience

TENEN ‘to tiptoe’
not salient (prop. mwe = 0.876)
N = 178

DUIM ‘thumb’
salient (prop. mwe = 0)
N = 209

Diversity in Language, Culture & Cognition Colloquium series, 09.11.2017
TENEN ‘TO TIPTOE’ (N >5)

- op zijn tenen lopen (99)
- op zijn tippen lopen (23)
- op zijn tenen gaan (8)
- op zijn voorvoeten lopen (8)
- trippelen (8)
Houdens van ‘to love’

Affect-sensitive (prop not-neutral = 1)

N = 214

Ademen ‘to breathe’

Neutral (prop = 0)

N = 104
HOUDEN VAN ‘to love’ (N >5)

- veel houden van (75)
- houden van (51)
- graag zien (26)
- veel houden af (8)
- veel ophebben met (7)
interim summary

- concept features influence lexical diversity
  - across semantic fields
  - across dialect areas
  → concepts that are more salient, less vague and not prone to affect show significantly less variation

- shortcoming: concept features can differ per person
  → social/lectal variation
  cf. salience of cacti: laymen vs. botanists
2. VARIATION IN CONCEPT FEATURES

Do experience-based features also correlate with lexical diversity?

Franco & Geeraerts (forthcoming)
examining variation in salience

- case-study 1: salience calculated using properties of the data → degree of salience across the data

- linguistic frequency
  *how often does a language user encounter a specific form to refer to the concept?*

  but: geographical stratification?

- experience with the concept/referent
  *how often does a language user encounter the concept?*

  → referential data to gauge (variation in) experiential salience

  cf. Anischanka et al. (2014), Geeraerts et al. (1994)

  cf. Geeraerts (2016)
Cirsium oleraceum (L.) Scop.  Moesdistel

Paul Van den Breemt

<table>
<thead>
<tr>
<th>Ecoregio</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duimen</td>
<td>2.4</td>
</tr>
<tr>
<td>Polder</td>
<td>0.4</td>
</tr>
<tr>
<td>Zand- en Zandleemstreek</td>
<td>12.2</td>
</tr>
<tr>
<td>Leemstreek</td>
<td>39.5</td>
</tr>
<tr>
<td>Kempen</td>
<td>7.2</td>
</tr>
<tr>
<td>Maasvallei</td>
<td>0</td>
</tr>
</tbody>
</table>

Rode Lijst  nb
Trendindex  -0.07
KFK         7
is variation in the amount of **lexical diversity** per plant related to the **referential frequency** of a plant?
DATA & METHODS
referential data

plant frequency measured on two levels:

- **global frequency:**
  - in Dutch-speaking Belgium as a whole
  - three measures (highly correlated)
  - absolute number of locations where plant occurs

- **local frequency:**
  - per ecological region
  - proportion of locations where plant occurs

Van Landuyt et al. (2006)
ecological regions
linguistic data

- semantic field of plants from WBD, WLD & WVD

(Vakgroep Nederlandse taalkunde UGent & Variaties vzw 2007)

- total number of investigated plants: N = 137

Diversity in Language, Culture & Cognition Colloquium series, 09.11.2017
calculating lexical diversity

- calculated per plant per ecological region

- type-token ratio (TTR)
  - number of different types / number of different tokens
  - higher value = more lexical diversity

- internal uniformity
  - $I_Z(Y) = \Sigma_{i=1}^{n} F_{Z,Y}(x_i)^2$
  - relative contribution of lexical item’s frequency to onomasiological profile
  - lower value = more lexical diversity
internal uniformity (I)

vergeet-mij-niet(je): 93.55% (N = 232)

blauwe kanne: 0.8% (N = 2)

onzevrouwetraantjes: 0.8% (N = 2)

... (8 lexemes with N = 2)

\[ I = 0.9355^2 + 8 \times (0.008^2) \]
\[ = 0.8757 \]
internal uniformity ($I$)

vergeet-mij-niet(je): 93.55% ($N = 232$)

blauwe kanne: 0.8% ($N = 2$)

onzevrouwetraantjes: 0.8% ($N = 2$)

... (8 lexemes with $N = 2$)

$$I = 0.9355^2 + 8 \times (0.008^2)$$

$$= 0.8757$$

---

den: 62.5% ($N = 10$)

grove den: 6.25% ($N = 1$)

mast: 31.25% ($N = 5$)

$$I = 0.625^2 + 0.0625^2 + 0.3125^2$$

$$= 0.4922$$
combining the referential and linguistic data

<table>
<thead>
<tr>
<th>concept</th>
<th>ecoregion</th>
<th>referential frequency</th>
<th>linguistic diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>global</td>
<td>global</td>
<td>local</td>
</tr>
<tr>
<td></td>
<td>frequency</td>
<td>frequency</td>
<td>frequency</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>beech</td>
<td>Campine</td>
<td>2229</td>
<td>248</td>
</tr>
<tr>
<td>beech</td>
<td>Dunes</td>
<td>2229</td>
<td>248</td>
</tr>
<tr>
<td>beech</td>
<td>Loamy</td>
<td>2229</td>
<td>248</td>
</tr>
<tr>
<td>beech</td>
<td>Polder</td>
<td>2229</td>
<td>248</td>
</tr>
<tr>
<td>beech</td>
<td>Sand-loamy</td>
<td>2229</td>
<td>248</td>
</tr>
</tbody>
</table>
combining the referential and linguistic data

<table>
<thead>
<tr>
<th>concept</th>
<th>referential frequency</th>
<th>linguistic diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>global frequency 1</td>
<td>global frequency 2</td>
</tr>
<tr>
<td>plant ecoregion</td>
<td>2229 248 678</td>
<td>2229 248 678</td>
</tr>
<tr>
<td>beech Campine</td>
<td>2229 248 678</td>
<td>2229 248 678</td>
</tr>
<tr>
<td>beech Dunes</td>
<td>2229 248 678</td>
<td>2229 248 678</td>
</tr>
<tr>
<td>beech Loamy</td>
<td>2229 248 678</td>
<td>2229 248 678</td>
</tr>
<tr>
<td>beech Polder</td>
<td>2229 248 678</td>
<td>2229 248 678</td>
</tr>
<tr>
<td>beech Sand-loamy</td>
<td>2229 248 678</td>
<td>2229 248 678</td>
</tr>
</tbody>
</table>

north of Belgium as a whole
per ecological region
methods & expectation

negative correlation plant frequency & lexical variation:
  • spearman rank correlation tests & correlation coefficients
  • global vs. local frequency: linear mixed-effects regression

→ TTR: negative correlations expected
  internal uniformity: positive correlations expected
RESULTS
p < 0.001 (spearman)
<table>
<thead>
<tr>
<th>local frequency (per ecological region)</th>
<th>global frequency (north of Belgium)</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequent</td>
<td>globally &amp; locally frequent</td>
</tr>
<tr>
<td>infrequent</td>
<td>globally infrequent</td>
</tr>
<tr>
<td>infrequent</td>
<td>locally infrequent</td>
</tr>
<tr>
<td>infrequent</td>
<td>globally &amp; locally infrequent</td>
</tr>
<tr>
<td>local frequency (per ecological region)</td>
<td>global frequency (north of Belgium)</td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>-------------------------------------</td>
</tr>
<tr>
<td>frequent</td>
<td>frequent (1) globally &amp; locally frequent</td>
</tr>
<tr>
<td>infrequent</td>
<td>(2) locally infrequent</td>
</tr>
</tbody>
</table>
hypothesis

(1) globally & locally frequent

(2) locally infrequent

(3) globally & locally infrequent

less lexical diversity

more lexical diversity

Diversity in Language, Culture & Cognition Colloquium series, 09.11.2017
global vs. local frequency

mean value for TTR and I per plant frequency category

expectation:
- TTR: increase → as expected
- I: increase → not as expected

stable in lmer, but small effect size
interim summary

• TTR: results as expected
  significant negative correlation between plant frequency &
  lexical variation
  → less frequent plants show more lexical diversity
• internal uniformity: results show opposite effect
  → names for frequent plants are not standardized enough
to be picked up by I

• why these diverging results?
• why small effect sizes?
<table>
<thead>
<tr>
<th>Plant (Ecological Region)</th>
<th>N Tokens</th>
<th>Distribution of Types</th>
<th>N Types</th>
<th>TTR</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Great Mullein, Loamy Region</td>
<td>26</td>
<td>lexeme_{1,18} occur once lexeme_{19,22} occur once</td>
<td>22</td>
<td>0.846</td>
<td>0.050</td>
</tr>
<tr>
<td>Bitter Dock, Polder Region</td>
<td>38</td>
<td>lexeme_{1,2} occur once lexeme_{3} occurs 3 times lexeme_{4} occurs 4 times lexeme_{5} occurs 10 times lexeme_{6} occurs 19 times</td>
<td>6</td>
<td>0.158</td>
<td>0.338</td>
</tr>
<tr>
<td>Black Locust, Sandy and Sand-Loamy Region</td>
<td>26</td>
<td>lexeme_{1,2,3} occur once lexeme_{4} occurs 23 times</td>
<td>4</td>
<td>0.154</td>
<td>0.787</td>
</tr>
<tr>
<td>Forget-Me-Not, Dunes Region</td>
<td>52</td>
<td>lexeme_{1} occurs 52 times</td>
<td>1</td>
<td>0.019</td>
<td>1</td>
</tr>
</tbody>
</table>
## TTR vs. I

<table>
<thead>
<tr>
<th>plant (ecological region)</th>
<th>N tokens</th>
<th>distribution of types</th>
<th>N types</th>
<th>TTR</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>great mullein, Loamy region</td>
<td>26</td>
<td>lexeme\textsubscript{1...18} occur once lexeme\textsubscript{19...22} occur once</td>
<td>22</td>
<td>0.846</td>
<td>0.050</td>
</tr>
<tr>
<td>bitter dock, Polder region</td>
<td>38</td>
<td>lexeme\textsubscript{1,2} occur once lexeme\textsubscript{3} occurs 3 times lexeme\textsubscript{4} occurs 4 times lexeme\textsubscript{5} occurs 10 times lexeme\textsubscript{6} occurs 19 times</td>
<td>6</td>
<td>0.158</td>
<td>0.338</td>
</tr>
<tr>
<td>black locust, Sandy and sand-loamy region</td>
<td>26</td>
<td>lexeme\textsubscript{1,2,3} occur once lexeme\textsubscript{4} occurs 23 times</td>
<td>4</td>
<td>0.154</td>
<td>0.787</td>
</tr>
<tr>
<td>forget-me-not, Dunes region</td>
<td>52</td>
<td>lexeme\textsubscript{1} occurs 52 times</td>
<td>1</td>
<td>0.019</td>
<td>1</td>
</tr>
</tbody>
</table>

Diversity in Language, Culture & Cognition Colloquium series, 09.11.2017
Daan (1969)

9: West-Flemish & Zeelandic Flemish

10: intermediate dialects between West- and East-Flemish

11: East-Flemish

15: Brabantic
other measures of experiential salience

usefulness of plant
e.g. lime tree:
- very infrequent
- little lexical diversity
→ used for artefacts

poisonousness of plant
e.g. black nightshade:
- very frequent
- a lot of lexical variation
→ negative connotation

etc.
other measures of experiential salience

the more experientially salient the plant, the less lexical diversity

→ less diversity for plants that...

- are useful
  - higher edibility rating (pfaf.org)
  - higher medicinal rating (pfaf.org)

- are poisonous (vs. not poisonous; data U Cornell)
results: TTR in WVD

- **edible** plants show significantly less lexical diversity \( (p < 0.01, \text{Adj R}^2: 0.065) \)

- plants that are useful for **medicinal applications** show significantly less lexical diversity \( (p < 0.05, \text{Adj R}^2: 0.039) \)

- **poisonousness**: NS, but on average, poisonous plants show more lexical diversity

- similar trends in WLD & WBD, but NS
experiential salience correlates negatively with lexical diversity (referential frequency, usefulness, poisonousness etc.)

experiential salience alone cannot account for variation
cf. small effect sizes

→ communicative relevance
CONCLUSIONS
conclusions

• **lexical diversity** occurs across languages, but also **within one language**

• dialectal lexical diversity correlates with **concept features**
  – across dialect areas and across semantic fields
  – concept features can be prone to **lectal variation**

• the amount of lexical diversity is higher for:
  – onomasiologically **vague** concepts
  – experientially and onomasiologically **less salient** concepts
  – concepts prone to **affect**
future research

• how does the lectal dimension interact with conceptual variation across languages?

• to what extent are these features relevant across language? same effect? vagueness & affect vs. salience?

• to what extent can these features account for diachronic variation? e.g. is the speed of lexical change slower for more salient concepts? (cf. Bochkarev et al. 2014)
Thank you!

Questions? Suggestions?

karlien.franco@kuleuven.be
www.karlienfranco.com
REFERENCES


Majid, A. & Burenhult, N. (2014). Odors are expressible in language, as long as you speak the right language. *Cognition*, 130(2), 266-270.


EXTRA
INTERACTIONS MODEL CASE-STUDY 1
interactions with semantic field

importance of concept characteristics differs per semantic field
lack of onomas. salience * onomas. vagueness

onomasiological vagueness has a larger effect for non-salient concepts
RESPONSE: CALCULATION
number of types per concept

e.g. TO GET MARRIED (TROUWEN): 3 types

- trouwen 181
- zich binden 1
- getrouwd worden 1

CORPULENT WOMAN (GROF GEBOUWDE VROUW): 131 types

- machochel 67
- schommel 41
- molenpaard 23
- machine 17
- kapitein 11
- mangel 11
- mokkel 8
- bai (fr.) 7
- madsel 5
- schokkel 5
- dikke madam 4
- ...
geographical fragmentation

- calculated as the proportion of dispersion and range

- dispersion: (weighted) average distance between the attestations of the unique words for a concept relative to other words for the same concept

- range: (weighted) average coverage of the words for a concept relative to the entire region where the concept occurs

(Geeraerts & Speelman 2010, Speelman & Geeraerts 2008)
<table>
<thead>
<tr>
<th>dispersion</th>
<th>range</th>
</tr>
</thead>
<tbody>
<tr>
<td>variants scattered across dialect area</td>
<td>each word type takes up almost entire dialect area</td>
</tr>
<tr>
<td>variants are found in nearby locations</td>
<td>each word type occurs in small geographical area</td>
</tr>
</tbody>
</table>

Dispersion & Range

is more scattered than
dispersion = 1.22

dispersion = 2.58

(between 1 and 4.401)
range = 0.82

range = 0.20

(proportion)
GEOGRAPHICAL VARIATION VS. NUMBER OF TYPES

Diversity in Language, Culture & Cognition Colloquium series, 09.11.2017
diverging effects: salience vs. vagueness & affect

“the results seem to imply that, while a higher degree of onomasiological vagueness and affect induce both more heterogeneous profiles for the concepts and, over and above geographical fragmentation, a larger amount of unique variants per concept, a higher lack of onomasiological salience only affects the former aspect of lexical diversity.”

(Franco 2017: 76)

- lack of onomasiological salience: hyperonymous, co-hyponymous, or possibly hyponymous names associated with more salient concepts

vs.

- affect-sensitive concepts: disposition to lexical creativity
- vague concepts: demarcational differences between speakers
model 1: number of types

- Vagueness -- lexical non-uniqueness
- Lack of salience -- prop. MWE's
- Affect -- prop. non-neutral ratings
- Lack of salience -- prevalence binary
- Semantic field
- Dictionary

Sum of squares

Diversity in Language, Culture & Cognition Colloquium series, 09.11.2017
model 2: weighted average dispersion

- vagueness -- lexical non-uniqueness
- lack of salience -- prop. MWE’s
- lack of salience -- prevalence binary
- affect -- prop. non-neutral ratings
- dictionary
- semantic field

Sum of squares

Diversity in Language, Culture & Cognition Colloquium series, 09.11.2017
model 3: weighted average range

- lack of salience -- prop. MWE's
- vagueness -- lexical non-uniqueness
- lack of salience -- prevalence binary
- dictionary
- semantic field
- affect -- prop. non-neutral ratings

Sum of squares

Diversity in Language, Culture & Cognition Colloquium series, 09.11.2017
<table>
<thead>
<tr>
<th>model term</th>
<th>estimate</th>
<th>SE</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>-0.4420</td>
<td>0.0360</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>dictionary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WBD</td>
<td>0.2120</td>
<td>0.0170</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>semantic field</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the house</td>
<td>0.0130</td>
<td>0.0460</td>
<td>NS</td>
</tr>
<tr>
<td>celebration &amp; entertainment</td>
<td>-0.0750</td>
<td>0.0460</td>
<td>NS</td>
</tr>
<tr>
<td>personality &amp; feelings</td>
<td>0.2060</td>
<td>0.0510</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>family &amp; sexuality</td>
<td>-0.0940</td>
<td>0.0700</td>
<td>NS</td>
</tr>
<tr>
<td>society, school &amp; education</td>
<td>-0.0730</td>
<td>0.0410</td>
<td>&lt; 0.1</td>
</tr>
<tr>
<td>lack of salience</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>proportion of hapaxes</td>
<td>2.3460</td>
<td>0.3630</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>vagueness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lexical non-uniqueness</td>
<td>0.0220</td>
<td>0.0020</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>affect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>proportion of non-neutral ratings</td>
<td>0.2070</td>
<td>0.0240</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>interaction terms</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sem. field (the house) : proportion of hapaxes</td>
<td>-1.2440</td>
<td>0.4940</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>sem. field (celebration &amp; entertainment) : proportion of hapaxes</td>
<td>-2.3410</td>
<td>0.4190</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>sem. field (personality &amp; feelings) : proportion of hapaxes</td>
<td>-1.8670</td>
<td>0.4040</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>sem. field (family &amp; sexuality) : proportion of hapaxes</td>
<td>1.2250</td>
<td>0.7220</td>
<td>&lt; 0.1</td>
</tr>
<tr>
<td>sem. field (society, school &amp; education) : proportion of hapaxes</td>
<td>-0.6110</td>
<td>0.4130</td>
<td>NS</td>
</tr>
<tr>
<td>sem. field (the house) : lexical non-uniqueness</td>
<td>-0.0020</td>
<td>0.0020</td>
<td>NS</td>
</tr>
<tr>
<td>sem. field (celebration &amp; entertainment) : lexical non-uniqueness</td>
<td>0.0000</td>
<td>0.0030</td>
<td>NS</td>
</tr>
<tr>
<td>sem. field (personality &amp; feelings) : lexical non-uniqueness</td>
<td>-0.0080</td>
<td>0.0020</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>sem. field (family &amp; sexuality) : lexical non-uniqueness</td>
<td>-0.0140</td>
<td>0.0060</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>sem. field (society, school &amp; education) : lexical non-uniqueness</td>
<td>-0.0060</td>
<td>0.0020</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>proportion of hapaxes : lexical non-uniqueness</td>
<td>-0.0570</td>
<td>0.0040</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>