A case-study on lexical variation in plant names using interlinked digitized dialect dictionaries

Karlien Franco, Barbara Piringer & Eveline Wandl-Vogt

ÖAW ACDH

Plant name variation

Quercus Robur ‘English oak’: little variation

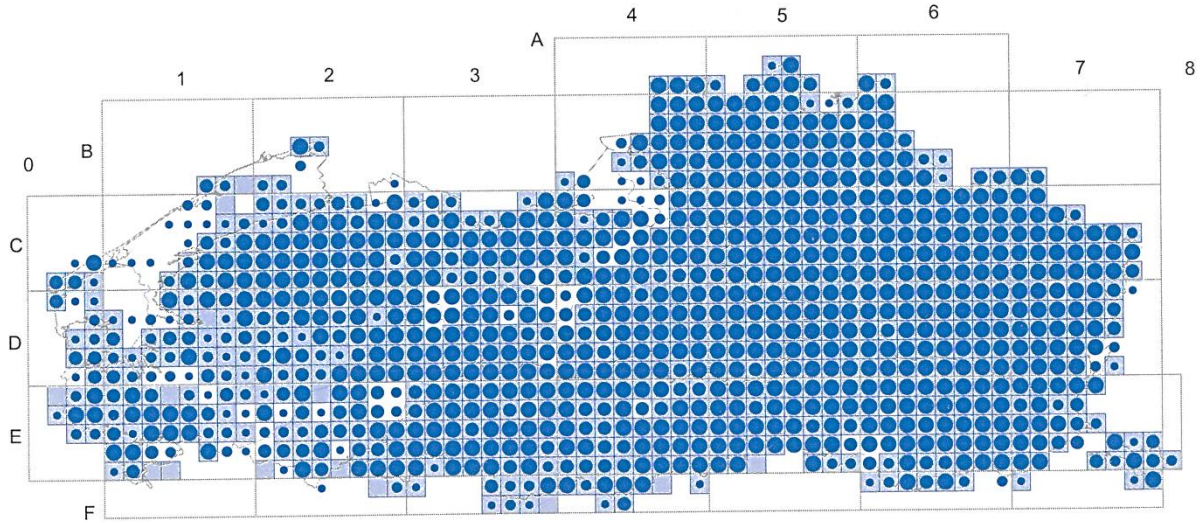
12 different Flemish dialectal names (6482 tokens)

e.g. *eik, eikelaar, kuipersboom, neikeboom, pestel*

occurs naturally throughout Flemish language area

Kaart Quercus Robur VL

Quercus robur L. **Zomereik** Paul Van den Bremt



Rode Lijst	nb
Trendindex	0.07
KFK	10

Ecoregio	**%**
Duinen	51.2
Polder	24.7
Zand- en Zandleemstreek	78.8
Leemstreek	87.4
Kempen	97.4
Maasvallei	59.2

Plant name variation

Quercus Robur ‘English oak’: little variation

12 different Flemish dialectal names (6482 tokens)

e.g. *eik, eikelaar, kuipersboom, neikeboom, pestel*

occurs naturally throughout Flemish language area

Primula Veris ‘cowslip’: a lot of variation

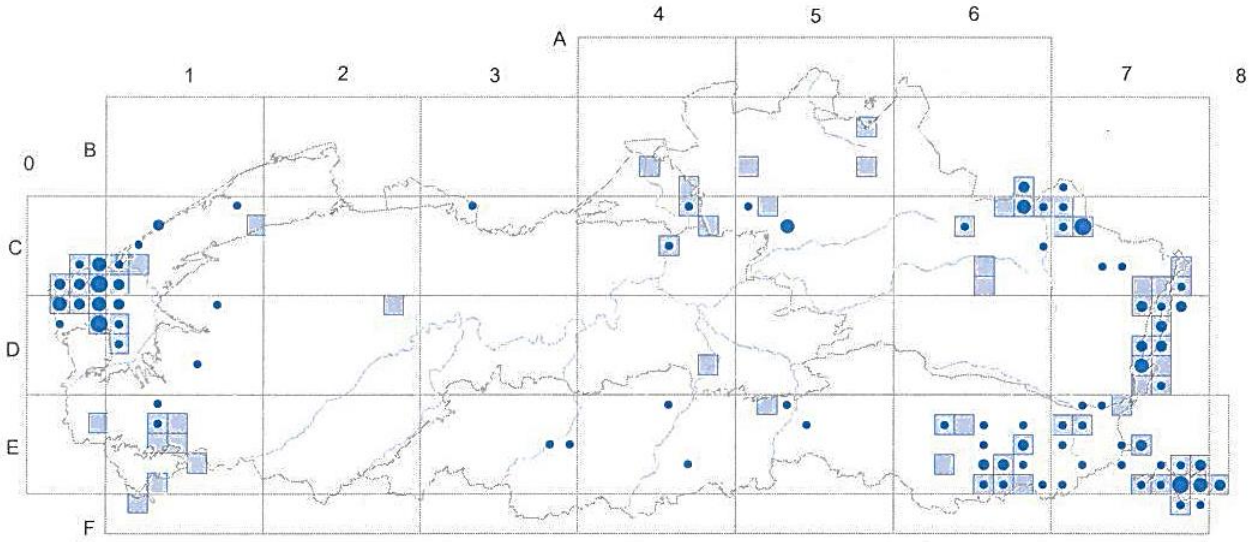
76 different dialectal names (523 tokens)

e.g. *bakbloem, eibloem(etje) , kerkesleutel , sleutelbloem(etje)*

does not grow frequently in Flemish language area

Kaart Primula Veris VL

Primula veris L. **Gulden sleutelbloem** Rein Brys



Rode Lijst	nb
Trendindex	-0.52
KFK	4

Ecoregio	**%**
Duinen	20.7
Polder	5.2
Zand- en Zandleemstreek	0.2
Leemstreek	3.3
Kempen	1.4
Maasvallei	19.7

Plant name variation

Quercus Robur ‘English oak’: little variation

12 different Flemish dialectal names (6482 tokens)

e.g. *eik, eikelaar, kuipersboom, neikeboom, pestel*

occurs naturally throughout Flemish language area

Primula Veris ‘cowslip’: a lot of variation

76 different dialectal names (523 tokens)

e.g. *bakbloem, eibloem(etje) , kerkesleutel , sleutelbloem(etje)*

does not grow frequently in Flemish language area

→ amount of variation in plant names correlates with referential plant frequency in Flemish dialect data

Plant name variation

Quercus Robur ‘English oak’: little variation

12 different Flemish dialectal names (6482 tokens)

e.g. *eik, eikelaar, kuipersboom, neikeboom, pestel*

occurs naturally throughout Flemish language area

Primula Veris ‘cowslip’: a lot of variation

76 different dialectal names (523 tokens)

e.g. *bakbloem, eibloem(etje) , kerkesleutel , sleutelbloem(etje)*

does not grow frequently in Flemish language area

→ amount of variation in plant names correlates with referential plant frequency in Flemish dialect data

Plant name variation

Quercus Robur ‘English oak’: little variation

12 different Flemish dialectal names (6482 tokens)

e.g. *eik, eikelaar, kuipersboom, neikeboom, pestel*

occurs naturally throughout Flemish language area

Primula Veris ‘cowslip’: a lot of variation

76 different dialectal names (523 tokens)

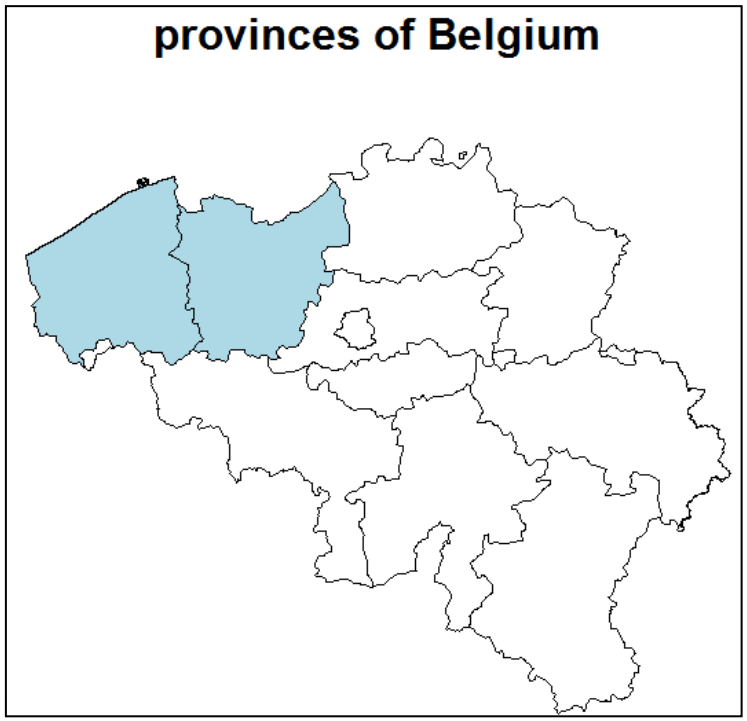
e.g. *bakbloem, eibloem(etje) , kerkesleutel , sleutelbloem(etje)*

does not grow frequently in Flemish language area

→ amount of variation in plant names correlates with referential plant frequency in Flemish dialect data

A pan-European perspective

- combining dialect dictionaries from two languages
  - dictionary of the Flemish dialects (WVD: dialects of Dutch in west of Flanders)
  - DBÖ (Bavarian Dialects of Austria)



Aim

- theoretical: further evidence for the relationship between plant familiarity and lexical variation

  → familiarity: operationalized as referential plant frequency

- practical:
  - to show that methodology used for Flemish data can be extended to a pan-European perspective
  - to discuss problems & perspectives for the future

Outline

- methodology
  1. interlinking the Bavarian and Flemish data
  2. adding measures of plant familiarity to the interlinked dataset

- analysis & results
  1. comparing lexical variation in the Bavarian and Flemish data
  2. correlating plant familiarity with lexical variation

- conclusions & implications for a pan-European perspective

1. interlinking the Bavarian and Flemish data

1. interlinking the Bavarian and Flemish data

- STEP 1: for both source datasets:

one-line-per-location  one-line-per-plant

1. interlinking the Bavarian and Flemish data

	G	H	K	O
1	plant	scientific nar	lexical variant	location
62121	sleutelbloem	primula veris	bakbloem	NA
62122	sleutelbloem	primula veris	bakbloem	Schoonaarde
62123	sleutelbloem	primula veris	bakkerinnetjes	NA
62124	sleutelbloem	primula veris	bakkers, bakkertjes	Oedelgem
62125	sleutelbloem	primula veris	bakkers, bakkertjes	NA
62126	sleutelbloem	primula veris	bakkers, bakkertjes	Donk
62127	sleutelbloem	primula veris	bakkers, bakkertjes	Brugge
62128	sleutelbloem	primula veris	bakkers, bakkertjes	Sint-Andries
62129	sleutelbloem	primula veris	bakkers, bakkertjes	Varsenare
62130	sleutelbloem	primula veris	bakkers, bakkertjes	Houtave
62131	sleutelbloem	primula veris	bakkers, bakkertjes	Oudenburg
62132	sleutelbloem	primula veris	bakkers, bakkertjes	Knokke
62133	sleutelbloem	primula veris	bakkers, bakkertjes	Beernem
62134	sleutelbloem	primula veris	bakkers, bakkertjes	Adegem
62135	sleutelbloem	primula veris	bakkers, bakkertjes	Reninge
62136	sleutelbloem	primula veris	bakkers, bakkertjes	Loppem
62137	sleutelbloem	primula veris	bakkers, bakkertjes	Brugge
62138	sleutelbloem	primula veris	bakkers, bakkertjes	Maldegem
62139	sleutelbloem	primula veris	bakkers, bakkertjes	Ruddervoorde
62140	sleutelbloem	primula veris	bakkers, bakkertjes	Stalhille

1. interlinking the Bavarian and Flemish data

	G	H	K	O
1	plant	scientific nar	lexical variant	location
62121	sleutelbloem	primula veris	bakbloem	NA
62122	sleutelbloem	primula veris	bakbloem	Schoonaarde
62123	sleutelbloem	primula veris	bakkerinnetjes	NA
62124	sleutelbloem	primula veris	bakkers, bakkertjes	Oedelgem
62125	sleutelbloem	primula veris	bakkers, bakkertjes	NA
62126	sleutelbloem	primula veris	bakkers, bakkertjes	Donk
62127	sleutelbloem	primula veris	bakkers, bakkertjes	Brugge
62128	sleutelbloem	primula veris	bakkers, bakkertjes	Sint-Andries
62129	sleutelbloem	primula veris	bakkers, bakkertjes	Varsenare
62130	sleutelbloem	primula veris	bakkers, bakkertjes	Houtave
62131	sleutelbloem	primula veris	bakkers, b	
62132	sleutelbloem	primula veris	bakkers, b	
62133	sleutelbloem	primula veris	bakkers, b	
62134	sleutelbloem	primula veris	bakkers, b	
62135	sleutelbloem	primula veris	bakkers, b	
62136	sleutelbloem	primula veris	bakkers, b	
62137	sleutelbloem	primula veris	bakkers, b	
62138	sleutelbloem	primula veris	bakkers, b	
62139	sleutelbloem	primula veris	bakkers, b	
62140	sleutelbloem	primula veris	bakkers, b	

	A	B	C	H	I	L
1	scientific_name	plant	global_freq	nr_types_wvd	nr_tokens_wvd	ttr_wvd
130	polygonum aviculare	varkensgr	971	6	55	0.1090909
131	polypodium vulgare	eikvaren	214	2	4	0.5
132	populus alba	witte abe	675	14	73	0.1917808
133	populus alba	populier (	675	34	315	0.1079365
139	primula veris	sleutelblc	84	76	523	0.1453155
143	prunus spinosa	sleedoorn	759	38	129	0.2945736
144	prunus spinosa	sleepruim	759	24	76	0.3157895
147	quercus robur	eik	920	12	6482	0.0018513
155	robinia pseudoacacia	robinia	732	4	31	0.1290323
157	rosa corymbifera	rozenbott	231	14	18	0.7777778
158	rubus fruticosus	braambes	851	33	811	0.0406905
159	rubus fruticosus	braamstru	851	71	668	0.1062874
163	rumex aquaticus	paardezur	NA	9	19	0.4736842
164	rumex obtusifolius	zuring (alg	972	17	131	0.129771

1. interlinking the Bavarian and Flemish data

- STEP 1: for both source datasets:

one-line-per-location → one-line-per-plant

→ one-line-per-plant datasets contain information about amount of lexical variation:

1. interlinking the Bavarian and Flemish data

- STEP 1: for both source datasets:

one-line-per-location → one-line-per-plant

→ one-line-per-plant datasets contain information about amount of lexical variation:

- number of types = number of different (unique) names
- number of tokens = number of records available per plant
- TTR measure = $\frac{number\ of\ types}{number\ of\ tokens}$
  - TTR = 1: a lot of variation
  - TTR = 0: no variation

1. interlinking the Bavarian and Flemish data

- STEP 2: link datasets by using scientific name:

scientific name	Dutch common name	nr types WVD	nr tokens WVD	TTR WVD	nr types DBÖ	nr tokens DBÖ	TTR DBÖ	German common names
agrostemma githago	bolderik	2	4	0.5	66	81	0.81	gew. kornrade, rade
anemone nemorosa	bosanemoon	51	261	0.20	87	100	0.87	busch-windröschen
...	...	...	...	...	...	...	...	...

1. problems with interlinking

- synonymy scientific names:

e.g. Crataegus (DBÖ) = Crataegus Monogyna (WVD)

→ manual corrections necessary

- only 36 plants occur in both datasets:

data from different regions: Alps versus region near North Sea

→ different ecological background & different plants

e.g. Primula Auricula: only occurs in the Alps and is very rare
→ not in Flemish dialect data

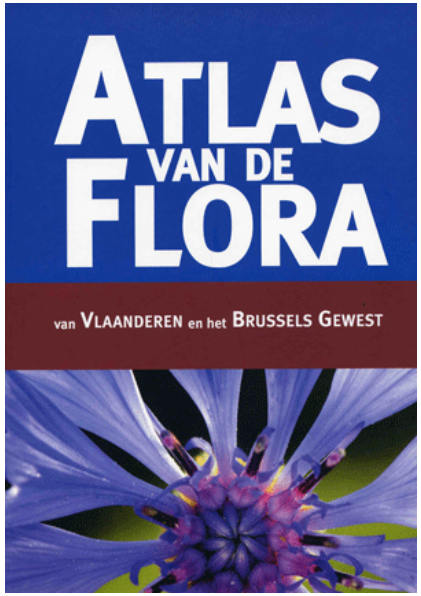
- variants of the same genus do occur

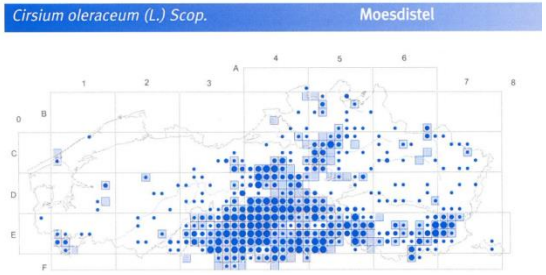
e.g. Anemone Hepatica only in DBÖ, Anemone Nemorosa in both
Arctium Lappa only in DBÖ, Arctium Minus only in WVD

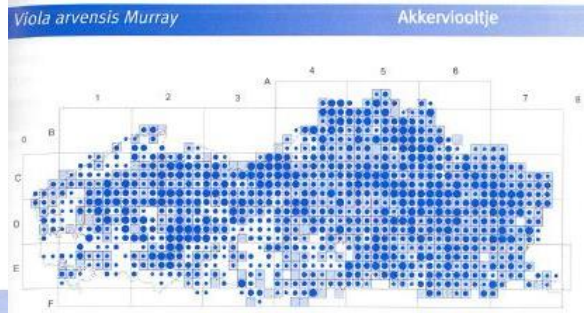
2. adding measures of plant familiarity

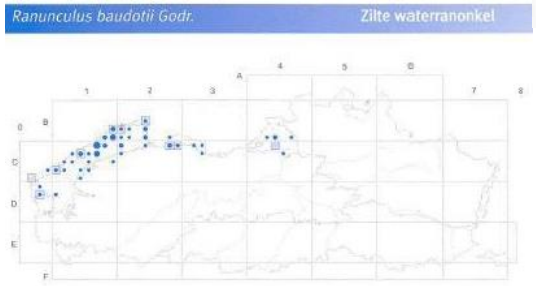
2. adding measures of plant familiarity

- Flemish data: *Atlas of the Flora of Flanders & Brussels* (Van Landuyt et al. 2006)
  - quantitative information about plant distribution: proportion of the area under investigation where plant occurs
  - database available online (http://flora.inbo.be)
  - on the basis of scientific name









2. adding measures of plant familiarity

- Bavarian data: no comparable plant distribution database freely available yet

- GBIF (Global Biodiversity Information Facility)?
  - http://www.gbif.org
  - huge international portal for collection of biological data
  - contains some comparable Austrian plant distribution data (U Wien), but only for 38 plants (not all in dataset)
  - occurrence counts in GBIF (human observation):
    - the more frequently a plant occurs in all the datasets of GBIF combined, the more well-known it is?
    - the opposite effect is possible too
  - search by scientific name (and synonyms)

2. GBIF example

10 results

[View results as map]

Configure Add a filter

BASIS OF RECORD	Human Observation ✖
LOCATION	-1.19 42.38...26.50 42.38 ✖ With NO known coordinate issues ✖
COUNTRY	Austria ✖
SCIENTIFIC NAME	Taraxacum officinale (L.) Weber ✖

	LOCATION	BASIS OF RECORD	DATE
237859480 · Cat. 282604 **Taraxacum officinale (L.) Weber** Published in Biosphärenpark Wienerwald - Pfaffstätten	Austria 48,02N, 16,25E	Human Observation	6 / 2009
165149424 · Cat. 208946 **Taraxacum officinale (L.) Weber** Published in Bernhardsthal	Austria 48,68N, 16,87E	Human Observation	6 / 2008

2. other measures that can gauge the familiarity of a plant

1. edibility rating
2. medicinal rating
   - Plants For A Future (http://pfaf.org)
   - over 7000 edible and medicinal plants
   - search by scientific name
   - 6-point scale (0-5)
   - hypothesis: edible and plants that are medically useful are more well-known → smaller amount of variation
3. poisonousness for humans & livestock
   - list published by Cornell U
   - binary: yes (= on Cornell U list) vs. no

   BUT not exhaustive: 39/208 plants included
   - hypothesis: poisonous plants are more well-known

   → smaller amount of variation

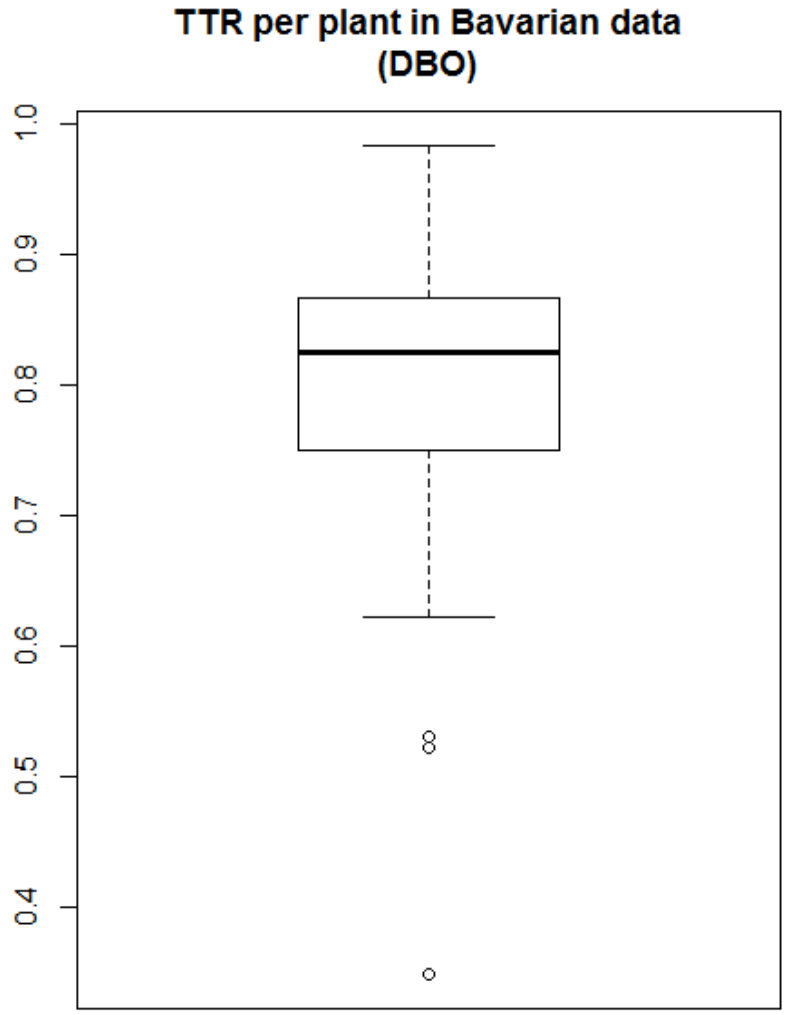
Outline

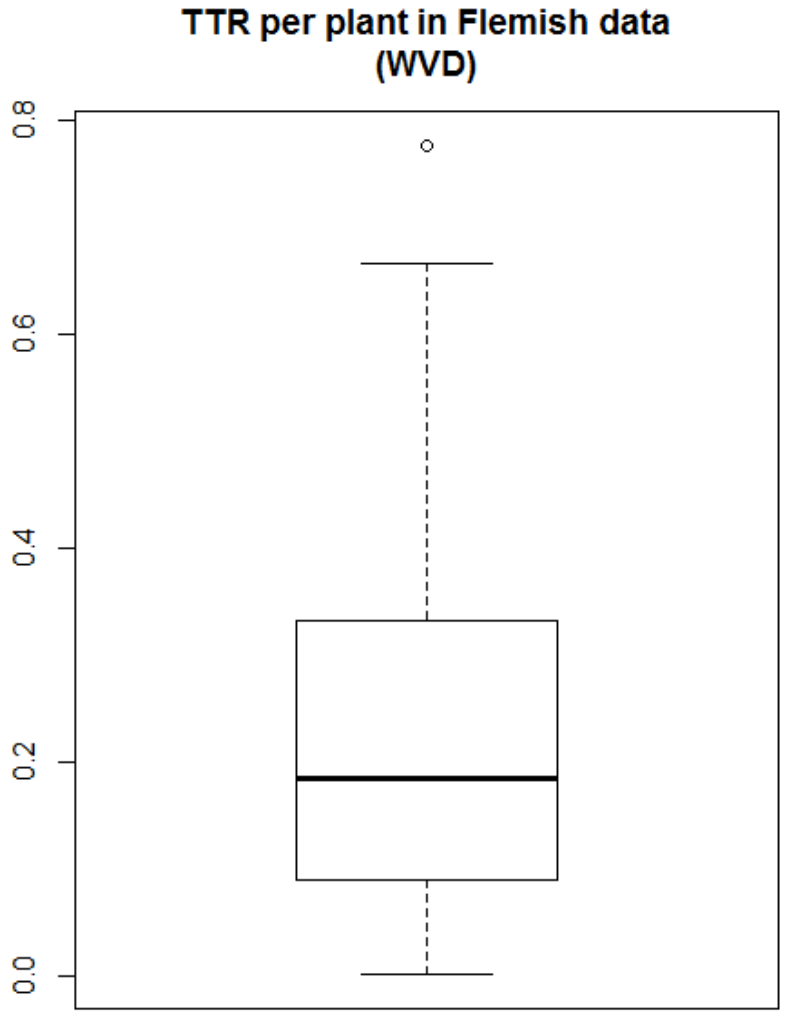
- methodology
  1. interlinking the Bavarian and Flemish data
  2. adding measures of plant familiarity to the interlinked dataset

- analysis & results
  1. comparing lexical variation in the Bavarian and Flemish data
  2. correlating plant familiarity with lexical variation

- conclusions & implications for a pan-European perspective

1. comparing lexical variation in the Bavarian and Flemish data





1. comparing lexical variation in the Bavarian and Flemish data

- more variation in Bavarian data

- possible explanations:
  - different sources: Flemish data based on questionnaires, but DBÖ-data from different sources (local dictionaries etc.)
  - TTR is sensitive to amount of data available:
    - is Flemish data more stable because of larger number of records per plant?
    - mean number of tokens per plant Flemish data: 322.8
    - mean number of tokens per plant Bavarian data: 102.22

2. correlating plant familiarity with lexical variation

- four measures of plant familiarity:
  - referential plant frequency (Atlas (Flemish), GBIF (Bavarian))
  - edibility rating
  - medicinal rating
  - poisonousness

- hypothesis: the more familiar the plant, the smaller the amount of lexical variation

  → more familiar = more referentially frequent
  higher edibility rating
  higher medicinal rating
  poisonous (vs. not poisonous)

2. correlating plant familiarity with lexical variation: results Flemish data

- referentially more frequent plants show a significantly smaller amount of lexical variation (spearman $p < 0.01$, $r = -0.310$)

- edible plants show a significantly smaller amount of lexical variation ($p < 0.01$, Adj R^2: 0.065)

- plants that are useful for medicinal applications show a significantly smaller amount of lexical variation ($p < 0.05$, Adj R^2: 0.039)

- the poisonousness of a plant does not have any significant effect, but on average, poisonous plants show more variation

poisonousness of a plant

e.g. black nightshade:

- very frequent
- a lot of lexical variation

→ dictionary can contain names that have to do with poisonousness of the berries:

duivelskersen ‘diabolical berries’, duivelskrallen ‘diabolical beads’, duivelskruid ‘diabolical herbs’, vergiftigde kersjes ‘poisonous cherries’, vergifbolletjes ‘poisonous balls’ etc.

2. correlating plant familiarity with lexical variation: results Bavarian data

- no significant effects
  - smaller amount of tokens per plant → results not as reliable?

- referential frequency shows opposite trend
  - but GBIF-data: not appropriate for our purposes?
  - maybe higher number of observations in GBIF of more rare plants mostly

- edibility, medicinal applications and poisonousness seem to show very weak trends in the same direction as results for Dutch data

  i.e. less variation for more useful plants,

  but more variation for poisonous plants

Outline

- methodology
  1. interlinking the Bavarian and Flemish data
  2. adding measures of plant familiarity to the interlinked dataset

- analysis & results
  1. comparing lexical variation in the Bavarian and Flemish data
  2. correlating plant familiarity with lexical variation

- conclusions & implications for a pan-European perspective

conclusions

- in the Flemish data, we find indications for the effect of plant familiarity (measured as referential frequency, edibility and medicinal usefulness) on the amount of lexical variation per plant
- we find similar, but non-significant weak trends in the Bavarian data

- we also find indications that more poisonous plants show more variation, but additional research is necessary

implications for the pan-European perspective

1. not all data is comparable

2. but comparing data from different countries and language regions offers new insights into the structure of the lexicon, the different backgrounds of the datasets and the culture of the countries

3. interlinked datasets can be analyzed by means of a single methodology, which reduces the amount of effort that is necessary

4. open-source is a must

for further information:

karlien.franco@kuleuven.be

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