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

Karlien Franco, Barbara Piringer & Eveline Wandl-Vogt
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

<table>
<thead>
<tr>
<th>Ecoregio</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duinen</td>
<td>51.2</td>
</tr>
<tr>
<td>Polder</td>
<td>24.7</td>
</tr>
<tr>
<td>Zand- en Zandleemstreek</td>
<td>78.8</td>
</tr>
<tr>
<td>Leemstreek</td>
<td>87.4</td>
</tr>
<tr>
<td>Kempen</td>
<td>97.4</td>
</tr>
<tr>
<td>Maasvallei</td>
<td>59.2</td>
</tr>
</tbody>
</table>

Rode Lijst: nb
Trendindex: 0.07
KFK: 10
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

<table>
<thead>
<tr>
<th>Ecoregio</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duinen</td>
<td>20.7</td>
</tr>
<tr>
<td>Polder</td>
<td>5.2</td>
</tr>
<tr>
<td>Zand- en Zandleemstreek</td>
<td>0.2</td>
</tr>
<tr>
<td>Leemstreek</td>
<td>3.3</td>
</tr>
<tr>
<td>Kempen</td>
<td>1.4</td>
</tr>
<tr>
<td>Maasvallei</td>
<td>19.7</td>
</tr>
</tbody>
</table>

Rode Lijst     nb
Trendindex     -0.52
KFK            4

ENeL Action meeting, Budapest 24.02.2017
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. bakbloom, 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

ENeL Action meeting, Budapest 24.02.2017
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

<table>
<thead>
<tr>
<th>G</th>
<th>H</th>
<th>K</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>plant</td>
<td>scientific nar</td>
<td>lexical variant</td>
<td>location</td>
</tr>
<tr>
<td>62121</td>
<td>sleutelbloem</td>
<td>primula veris</td>
<td>bakbloem</td>
</tr>
<tr>
<td>62122</td>
<td>sleutelbloem</td>
<td>primula veris</td>
<td>bakbloem</td>
</tr>
<tr>
<td>62123</td>
<td>sleutelbloem</td>
<td>primula veris</td>
<td>bakkerinnetjes</td>
</tr>
<tr>
<td>62124</td>
<td>sleutelbloem</td>
<td>primula veris</td>
<td>bakkers, bakkertjes</td>
</tr>
<tr>
<td>62125</td>
<td>sleutelbloem</td>
<td>primula veris</td>
<td>bakkers, bakkertjes</td>
</tr>
<tr>
<td>62126</td>
<td>sleutelbloem</td>
<td>primula veris</td>
<td>bakkers, bakkertjes</td>
</tr>
<tr>
<td>62127</td>
<td>sleutelbloem</td>
<td>primula veris</td>
<td>bakkers, bakkertjes</td>
</tr>
<tr>
<td>62128</td>
<td>sleutelbloem</td>
<td>primula veris</td>
<td>bakkers, bakkertjes</td>
</tr>
<tr>
<td>62129</td>
<td>sleutelbloem</td>
<td>primula veris</td>
<td>bakkers, bakkertjes</td>
</tr>
<tr>
<td>62130</td>
<td>sleutelbloem</td>
<td>primula veris</td>
<td>bakkers, bakkertjes</td>
</tr>
<tr>
<td>62131</td>
<td>sleutelbloem</td>
<td>primula veris</td>
<td>bakkers, bakkertjes</td>
</tr>
<tr>
<td>62132</td>
<td>sleutelbloem</td>
<td>primula veris</td>
<td>bakkers, bakkertjes</td>
</tr>
<tr>
<td>62133</td>
<td>sleutelbloem</td>
<td>primula veris</td>
<td>bakkers, bakkertjes</td>
</tr>
<tr>
<td>62134</td>
<td>sleutelbloem</td>
<td>primula veris</td>
<td>bakkers, bakkertjes</td>
</tr>
<tr>
<td>62135</td>
<td>sleutelbloem</td>
<td>primula veris</td>
<td>bakkers, bakkertjes</td>
</tr>
<tr>
<td>62136</td>
<td>sleutelbloem</td>
<td>primula veris</td>
<td>bakkers, bakkertjes</td>
</tr>
<tr>
<td>62137</td>
<td>sleutelbloem</td>
<td>primula veris</td>
<td>bakkers, bakkertjes</td>
</tr>
<tr>
<td>62138</td>
<td>sleutelbloem</td>
<td>primula veris</td>
<td>bakkers, bakkertjes</td>
</tr>
<tr>
<td>62139</td>
<td>sleutelbloem</td>
<td>primula veris</td>
<td>bakkers, bakkertjes</td>
</tr>
<tr>
<td>62140</td>
<td>sleutelbloem</td>
<td>primula veris</td>
<td>bakkers, bakkertjes</td>
</tr>
</tbody>
</table>
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

→ 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 = \( \frac{\text{number of types}}{\text{number of tokens}} \)
  - 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:

<table>
<thead>
<tr>
<th>scientific name</th>
<th>Dutch common name</th>
<th>nr types WVD</th>
<th>nr tokens WVD</th>
<th>TTR WVD</th>
<th>nr types DBÖ</th>
<th>nr tokens DBÖ</th>
<th>TTR DBÖ</th>
<th>German common names</th>
</tr>
</thead>
<tbody>
<tr>
<td>agrostemma githago</td>
<td>bolderik</td>
<td>2</td>
<td>4</td>
<td>0.5</td>
<td>66</td>
<td>81</td>
<td>0.81</td>
<td>gew. kornrade, rade</td>
</tr>
<tr>
<td>anemone nemorosa</td>
<td>bosanemoom</td>
<td>51</td>
<td>261</td>
<td>0.20</td>
<td>87</td>
<td>100</td>
<td>0.87</td>
<td>buschwindröschen</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
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

  - 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

<table>
<thead>
<tr>
<th>BASIS OF RECORD</th>
<th>Human Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOCATION</td>
<td>1.19 42.38...26.50 42.38</td>
</tr>
<tr>
<td>COUNTRY</td>
<td>Austria</td>
</tr>
<tr>
<td>SCIENTIFIC NAME</td>
<td>Taraxacum officinale (L.) Weber</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LOCATION</th>
<th>BASIS OF RECORD</th>
<th>DATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>Human Observation</td>
<td>6 / 2009</td>
</tr>
<tr>
<td>Austria</td>
<td>Human Observation</td>
<td>6 / 2008</td>
</tr>
</tbody>
</table>

Published in Biosphärenpark Wienenwald · Paffstäten
Published in Bacharstedt
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

![Box plots comparing TTR per plant in Bavarian and Flemish data](image-url)
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
  \[ \text{more familiar} = \text{more referentially frequent} \]
  \[ \text{higher edibility rating} \]
  \[ \text{higher medicinal rating} \]
  \[ \text{poisonous (vs. not poisonous)} \]
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