DiaHClust: an iterative hierarchical clustering approach for identifying stages in language change

Christin Schätzle[1] and Hannah Booth[2]
[1]University of Konstanz, [2]Ghent University

1st International Workshop on Computational Approaches to Historical Language Change 2019

Motivation

- Historical linguistic investigations often make use of pre-determined periodization schemes to assess the progress of a change
- Pre-determined periodization schemes can be problematic:
  - little connection to the linguistic phenomenon under investigation
  - language-external influences (e.g., historical milestones)
  - often designed to be equidistant
- True trajectory of a change might be concealed, transitional periods obscured

⟶ **Solution:** Data-driven methods for the identification of stages in language change

Example – Historical English

Old English: c. 700-1100
| ← **1066: Norman conquest**
Middle English: c. 1100-1500
| ← **1476: Arrival of printing**
Early Modern English: c. 1500-1700

VNC

- Variability-based Neighbor Clustering (Gries and Hilpert 2008)
- Hierarchical clustering approach which is sensitive to the temporal ordering of data
  - Data with similar linguistic characteristics in the same cluster (i.e., period)
  - Breaks between periods at points where the characteristics of the data show a quantifiable shift
- Developed to assess how individual linguistic features change across different contexts (distributional properties)
- Our implementation of the VNC approach:

Algorithm 1 Implementation of VNC

```
function VNC
  ▷ Manipulation of distance matrix (dist):
    for i = 1 to numberOfRows(dist) do
      for j = 1 to i do
        if not i = j then
          dist[i, j] = max(dist)
  ▷ Clustering process:
    for k = 1 to numberOfRows(dist) do
      find m, n for which dist[m, n] = min(dist)
      dist[, n] = (dist[, n] + dist[, n+1]) / 2
      if dist[1, 1] = min(dist) then
        delete dist[1, ]
      else
        dist[m, ] = (dist[m−1, ] + dist[m, ]) / 2
```

References

Hannah Booth, Christin Schätzle, Kersti Börjars, and Miriam Butt. 2017. Dative subjects and the rise of positional licensing in Icelandic. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG'17 Conference, University of Konstanz*, pages 104–124. CSLI Publications, Stanford, CA.

Aaron Ecay and Susan Pintzuk. 2016. The syntax of Old English poetry and the dating of Beowulf. In Leonard Neidorf et al., eds, *Old English Philology: Studies in Honour of R.D. Fulk*, pages 219–258. D.S Brewer, Cambridge.

Stefan Th. Gries and Martin Hilpert. 2008. The identification of stages in diachronic data: variability-based neighbour clustering. *Corpora*, 3(1):59–81.

Anthony Kroch. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change*, 1:199–244.

Joel C. Wallenberg, Anton Karl Ingason, Einar Freyr Sigurðsson, and Eiríkur Rögnvaldsson. 2011. *Icelandic Parsed Historical Corpus (IcePaHC)*, version 0.9.

Richard Zimmermann. 2014. Dating hitherto undated Old English texts based on text-internal criteria. *Ms.*, University of Geneva.

DiaHClust

- Data-driven periodization methodology developed for the identification of stages in syntactic change
- Syntactic change is deeply interactional (e.g., Kroch 1989) – assessing distributional properties of individual features is not sufficient for understanding these interactions
- We propose to use syntactic vectors to inform the periodization with existing knowledge about the language's syntactic system over time (cf. Zimmermann 2014, Ecay and Pintzuk 2016):

(1) Text A = {feature$_1$, feature$_2$, …, feature$_n$}
Text B = {feature$_1$, feature$_2$, …, feature$_n$}
⋮

Change	1150.FIRSTGRAMMAR	1150.HOMILIUBOK	1210.JARTEIN	1210.THORLAKUR
SbjDat	0.6098	2.9137	5.8154	4.0936
OvertExpl	0.0000	0.2749	0.0000	0.1949
V1	19.5122	12.0396	32.2377	30.9942
SbjPrefinite	39.6341	44.6399	45.5120	46.3938
VO	56.0000	54.9296	47.0588	34.4828
StylisticFronting	3.0488	2.5838	0.6321	0.9747

Example dataset showing texts as syntactic vectors

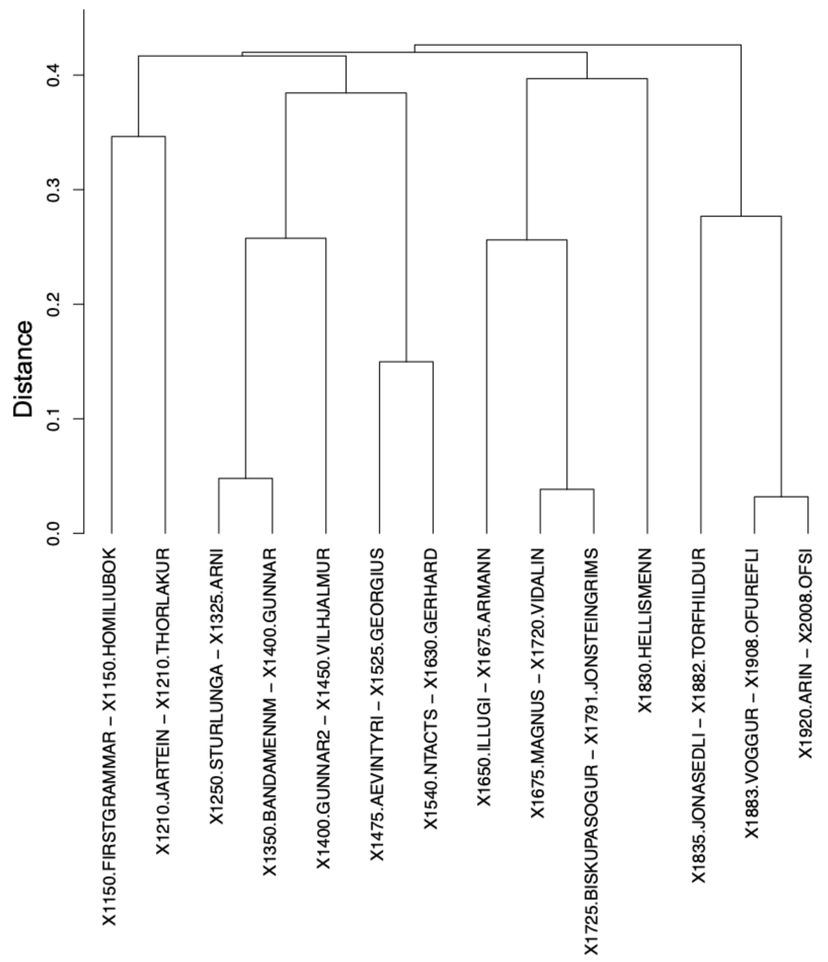
Algorithm 2 DiaHClust methodology

```
function DIAHCLUST
  repeat
    aggregate(data)
    dist = distanceMatrix(cor(data))
    clust = vnc(dist)
    plot(clust)
    computeOptimalClustering(clust)
  until numberOfClusters < 10
```

- DiaHClust is based on VNC, but employs an extra iterative layer of hierarchical clustering
  - This allows us to begin at text-level, tracing the clustering process until the final larger time stages are identified
  - We calculate silhouette coefficients ($s(i)$) to automatically identify the optimal clustering at each iteration which in turn informs the next clustering step
  - Information about the composition of clusters is given at each step of the iteration
- We implemented the DiaHClust methodology as an R package
  → Package and code available at https://github.com/christinschaetzle/diaHClust

Identifying stages in language change

- Case study on Icelandic – the most conservative Germanic language; investigation of syntactic developments which are known to interact
- Standard periodization for Icelandic influenced by language-external factors (first translation of the New Testament in the 16th century, equidistant periods)
- Extraction of changing features from IcePaHC (Wallenberg et al. 2011), a syntactically annotated corpus of historical Icelandic (1150-2008)
  → dative subjects, expletives, V1 declaratives, subject position, VO order, Stylistic Fronting



DiaHClust results for syntactic change in IcePaHC

- VNC clustering and calculation of silhouette values: 28 clusters for 61 IcePaHC texts
  → DiaHClust until number of clusters < 10
- **DiaHClust:**
  - 6 time stages → 1150–1210, 1250–1450, 1475–1630, 1830–1830, 1835–2008
  - '1830.HELLISMENN' identified as outlier
  - 5 time stages after outlier removal
    → 1150–1210, 1250–1450, 1475–1630, 1650–1882, 1883–2008
  - Average $s(i) > 0.5$ (coherent clustering)
  - Genre effect carved out: mainly religious texts in stage 1475–1630
- DiaHClust periodization provides insights into how IcePaHC texts behave with respect to syntactic phenomena
- DiaHClust reveals that syntactic change follows a more gradual trajectory in Icelandic than has been previously assumed

Traditional Periodization

Time periods	% V1	% DatSubj
1150-1349	20.6	3.9
1350-1549	19.9	3.2
1550-1749	14.8	3.7
1750-1899	18.4	3.8
1900-2008	2.7	5.8

⟹

DiaHClust

Time periods	% V1	% DatSubj
1150-1210	23.7	3.4
1250-1450	23.2	4.0
1475-1630	6.9	2.6
1650-1882	15.6	4.1
1883-2008	2.3	5.5

Proportion of V1-clauses/dative subjects in IcePaHC as per Booth et al. (2017) vs. DiaHClust periods