Detecting Regional and Age Variation in a Growing Corpus of DGS

Thomas Hanke, Reiner Konrad, Gabriele Langer, Arike Müller, Sabrina Wähl
University of Hamburg, Institute of German Sign Language and Communication of the Deaf

DGS Corpus (2009-2023)
Informants:
- Number of informants: 330
- Controlled sample: balanced for 13 regions (according to estimated size of deaf population), see map
- 4 age groups: 18-30, 31-45, 46-60, 61+
- no underage informants
- gender
- Native and near-native signers, rooted in the Deaf community, regionally rooted (>10 years in the same region)

Method:
- Filmed conversations and staged communicative events (Makino et al. 2010)
- Materials collected, lexicalised and accessible through iLex (Hanke/Storz 2008)

Data:
- Data collection: 2009-2012
- Natural signing in context
  - > 560 hours footage of relevant signing
- Lemmatised: 425,000 tokens (2017-07-18) = 65 hours
- Including 26,500 tokens of task (elicitation of isolated signs for some concepts with known high regional lexical variation such as signs for colours and months in this poster this only applies to the example FRAU ‘woman’)

Starting Point:
- Corpus data balanced, but the annotated part is not.
  - No rigid statistical measures available, but often enough tokens to detect interesting cases of age and regional variation.
  - Such detections may guide detailed annotation as we will not be able to annotate the whole corpus in detail.

- Here we only cover variation identifiable on the single-token level, i.e. lexical variation, but not syntactic or morphological variation and only certain aspects of phonological/phonetic variation.

- Focus on semantic clusters, i.e. groups of signs with roughly the same meaning
  - Lexical variation should take place within these clusters
  - Basic annotation limits identification of phonological/phonetic variation, detailed annotation only available for a very small part of the corpus.
  - Only clusters that as a whole have enough tokens from all age groups/regions are looked at. (Cut-offs determined empirically, with no claim for statistical relevance.)
  - Age: min. 15 tokens in cluster from all 13 regions
  - Region: min. 100 tokens in cluster from at least 26 informants

Procedure:
- Rank eligible clusters by a heterogeneity index computed from either standard deviation or linear regression on member signs distributions
  - Cluster heterogeneity defined as number of cluster members exceeding a heterogeneity index threshold
  - Linear regression favours some distribution patterns over others (age difference and action of geolinguistics)
  - Inspect candidates by visualising the data in age distribution charts or maps (see examples on the right)
  - For regional variation, this may require several steps removing dominant stable forms from the cluster (cf. FRAU ‘woman’).

Towards empirical evidence for dialect regions:
- We identify some recurrent patterns for regional distribution
  - The most prominent dialect region candidates are Bavaria, Saxony, Westphalia, East Germany.
  - In combination of regional and age distribution, we can observe some cases of variant spreading.

Towards empirical evidence for language change:
- We identify some recurrent patterns for language change based on the concept of aperiod time, widely used in sociolinguistics: A gradient age distribution in a language community indicates language change.
- This is based on the assumption that an individual’s vocabulary or vernacular is fixed at some point in their life span with no further significant or systematic change. Real-time studies confirmed in many cases the usefulness and validity of this concept and should be used as complementary methodologies (Bowie 2005, Sankoff 2006).
- The applicability of apparent time to fast-changing sign languages still needs validation through real-time studies!

Results:
- Efficient: A good part of the candidate clusters show variation.
  - The signs undergoing variation are earmarked for their lexicographic description.
  - Not so clear-cut cases and cases with comparably few token counts will get higher priority for future annotation.
- Plausible
  - The “usual suspects” are among the candidates if we have enough tokens for them.
  - Not at odds with DGS regional variation reported in the literature (tagging signs as “North”, “South”, “Bavaria” etc.).

Summary:
- We suggest a procedure to detect potential variation within a corpus with annotation still in progress. This procedure allows us to have a clearer look at specific signs to confirm variation.
- Second-order observations as language change and the identification of dialect regions are highly speculative at this point of time. Nevertheless, hypotheses generated can be used to guide deployment strategies for the DGS Feedback, our approach to involving the language community with more fine-grained data collections.

References
- Makino, Tomomi; Hanke, Thomas; Konrad, Reiner; Müller, Arike; Wähl, Sabrina; Mühlbauer, Gudrun; Kaiser, Marc; Tschöpe, Christoph; Young, Joanne (2010). Database Test for Investigating Sign Language Communities and Sign Language Community - A Case Study on Sign Languages in Germany. In International Journal of Language and Communication Disorders, Vol. 45, No. 2, pp. 110-133.
- Tschöpe, Christoph; Young, Joanne; Endres, Sabrina; Konrad, Reiner; Hanke, Thomas (2010). Real-time methods in the DGS Feedback: A case study to test the efficiency of the “apparent time” approach. In Frontiers in Biomedicine and Language: From Theory to Practice, pp. 102-111 (Eds. P. Sgall et al.). Berlin: Mouton de Gruyter.

Poster presented at the Workshop: Corpus-based approaches to sign language typology: Into the second decade, Birmingham, 24th July, 2017