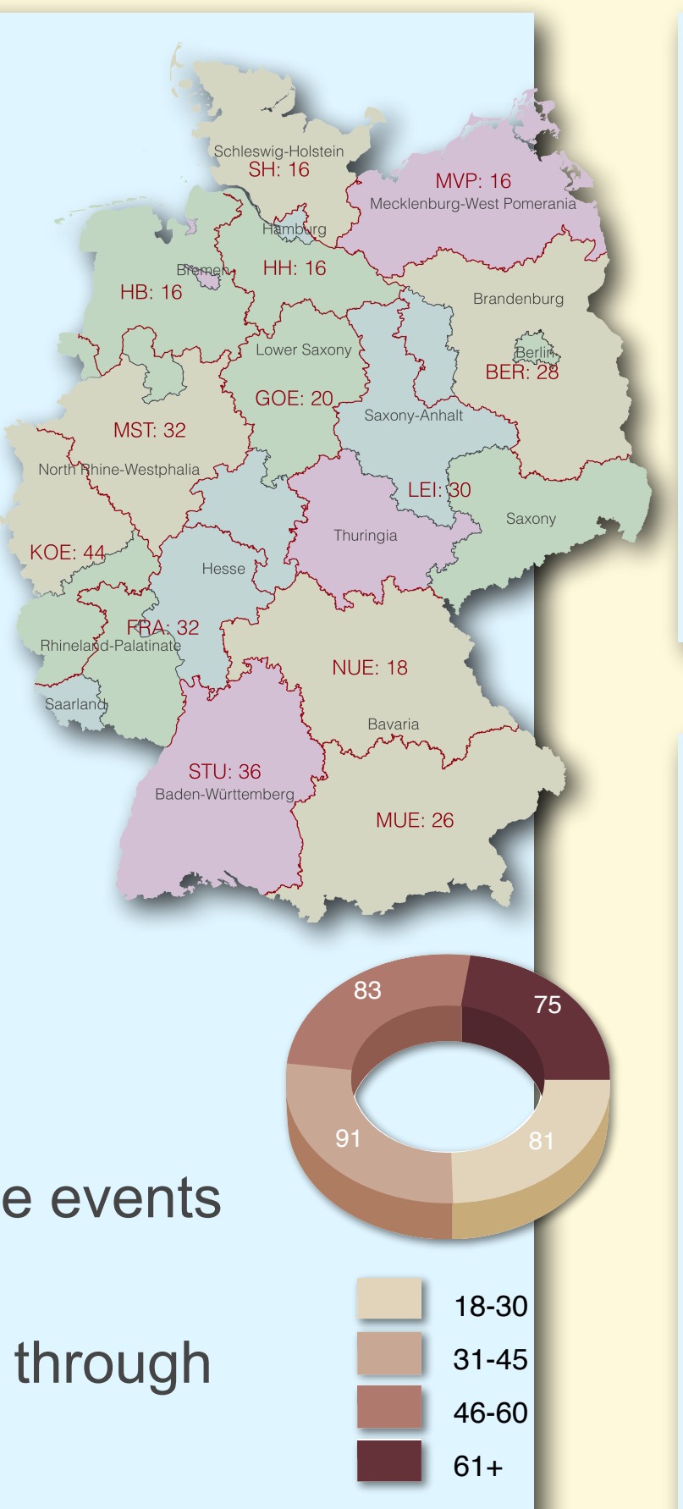
Detecting Regional and Age Variation in a Growing Corpus of DGS

Thomas Hanke, Reiner Konrad, Gabriele Langer, Anke 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 (Nishio et al. 2010)
- Multi-modal corpus, lemmatised 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 28,500 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 axion 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’).

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 closer 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.

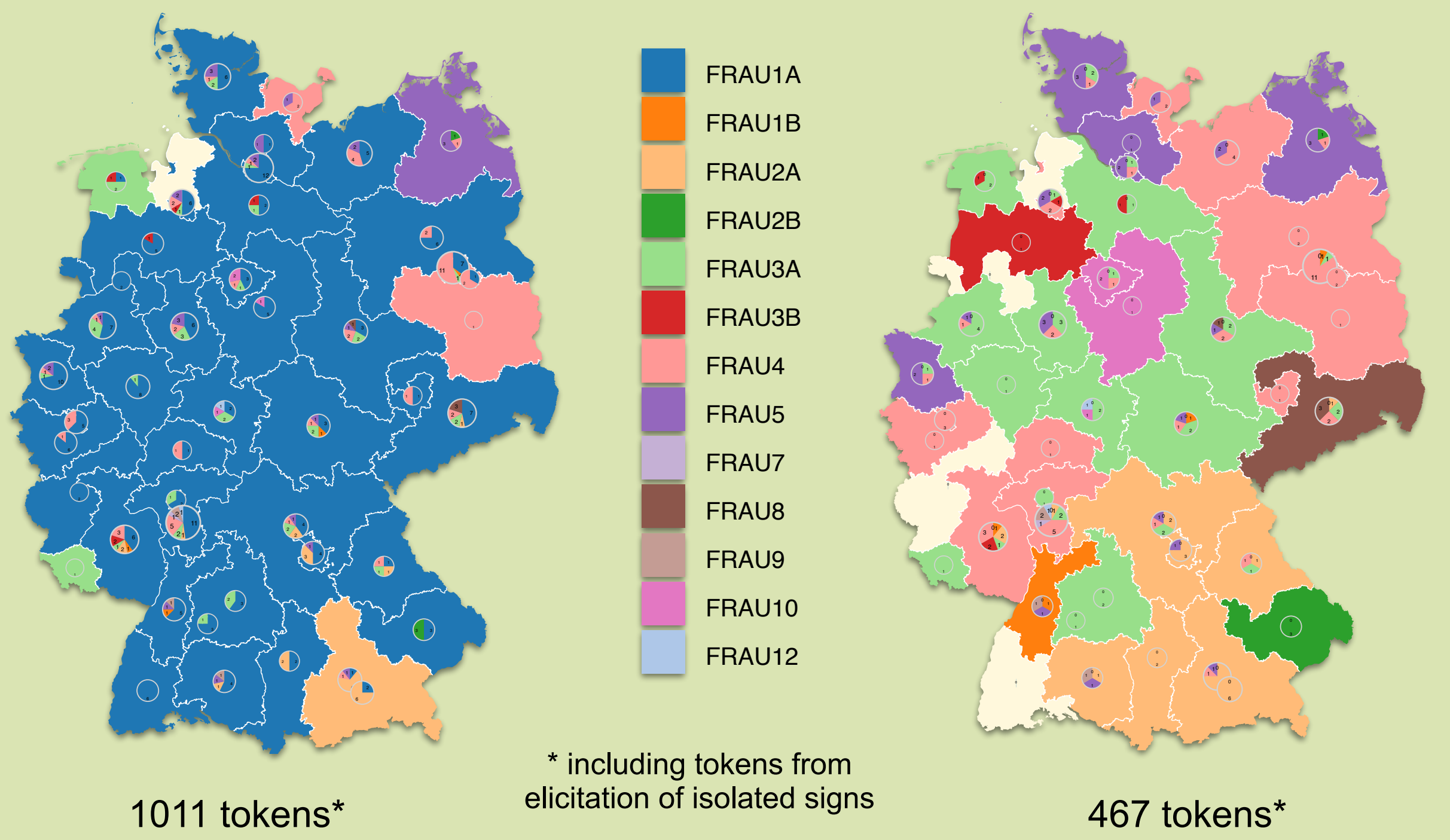
Towards empirical evidence for dialect regions

- 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 apparent 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!

FRAU ‘woman’: Removal of main variant FRAU1A from map reveals clearer distribution of lesser-used regional variants (see right map)

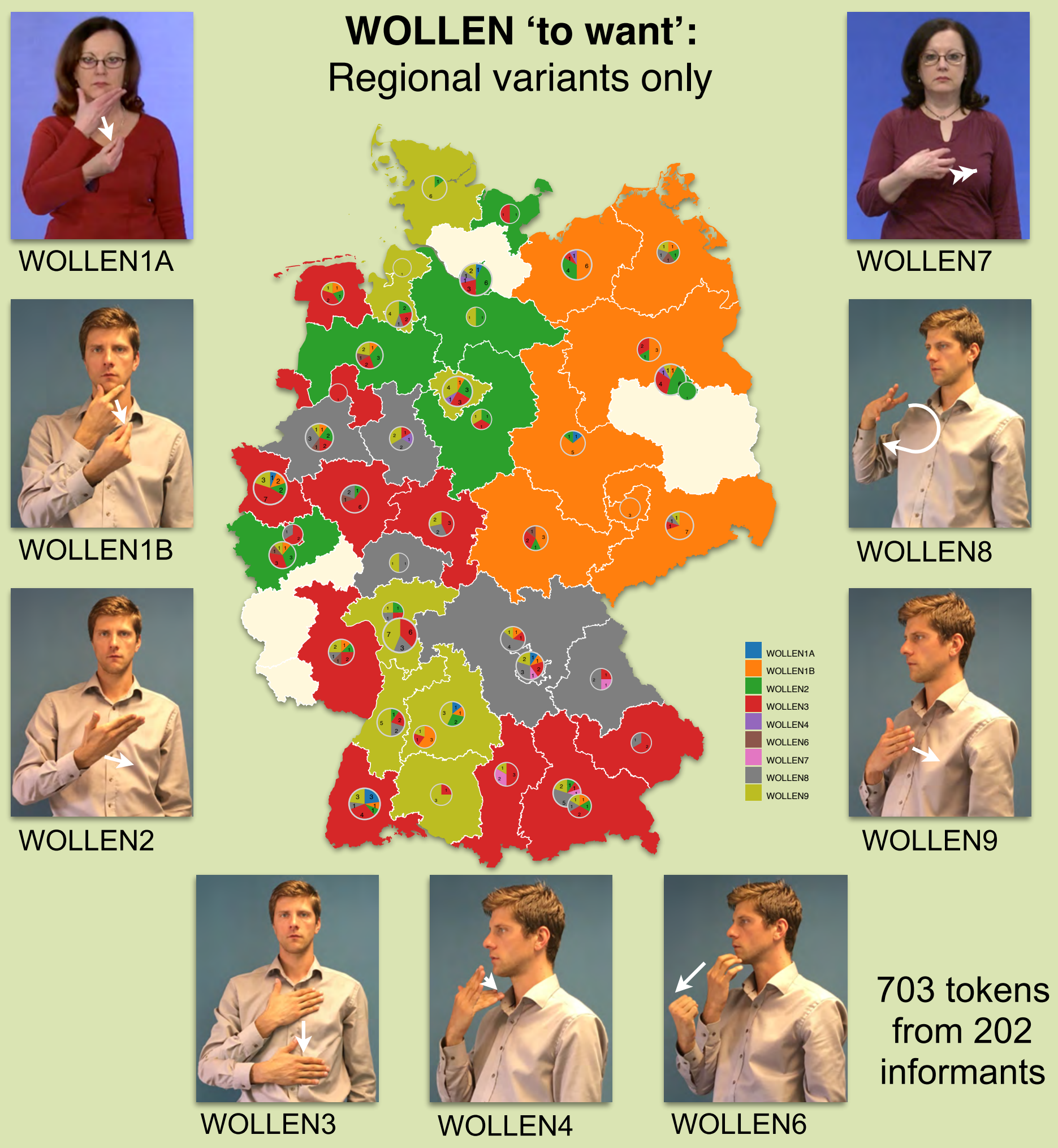


1011 tokens*
from 241 informants

* including tokens from elicitation of isolated signs

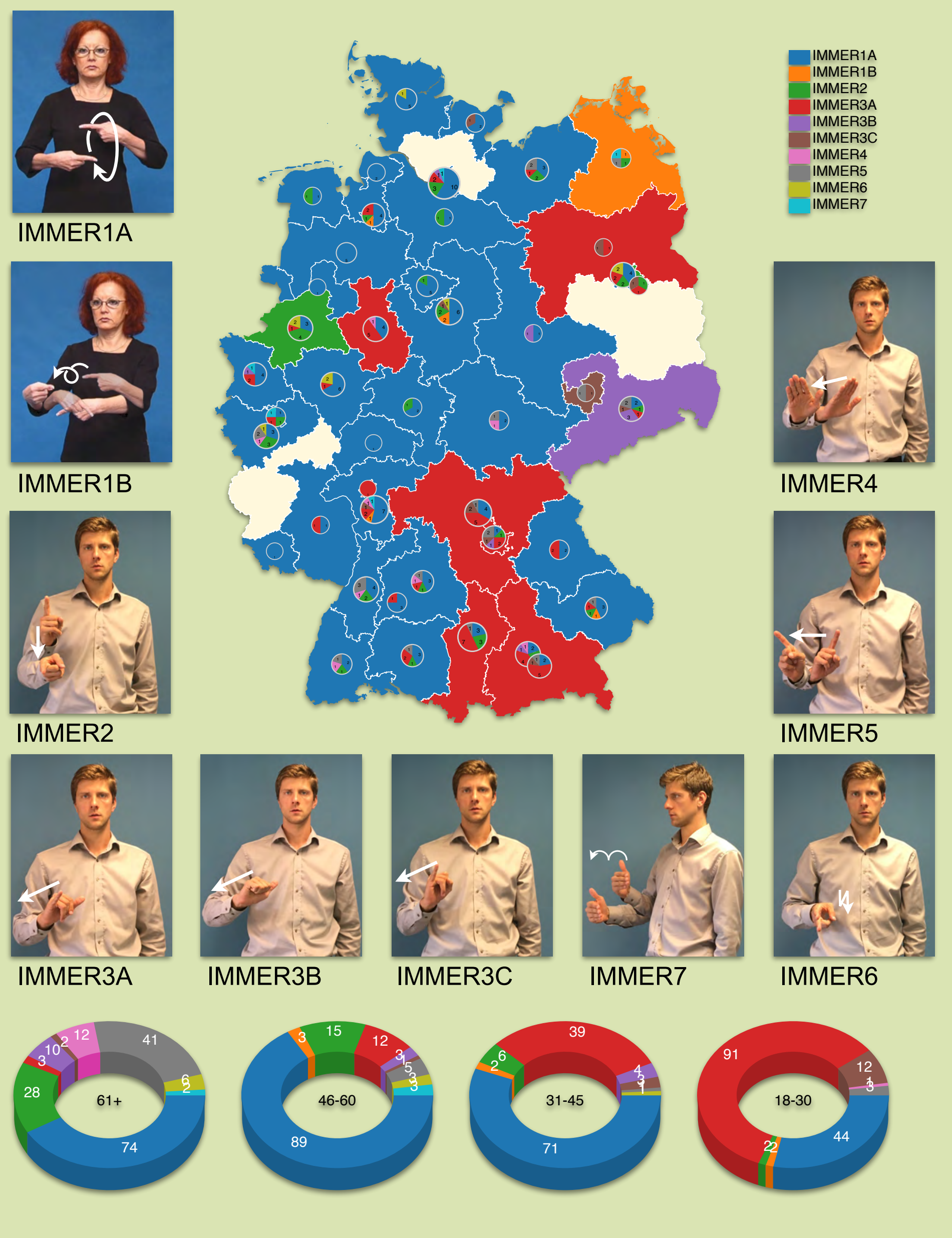
467 tokens*
from 137 informants

WOLLEN ‘to want’:
Regional variants only



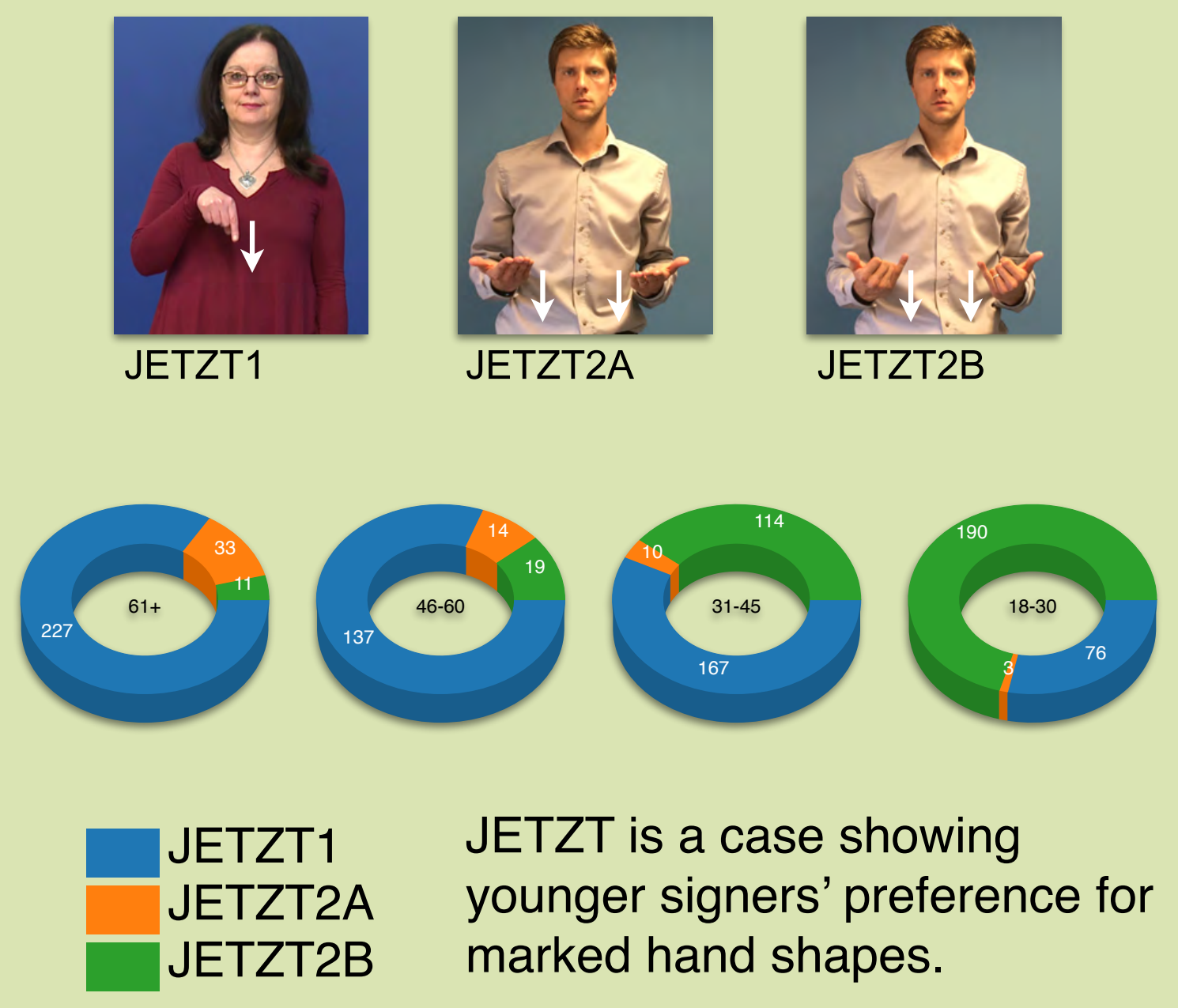
703 tokens
from 202
informants

IMMER ‘always’: A case of levelling?



594 tokens from 187 informants

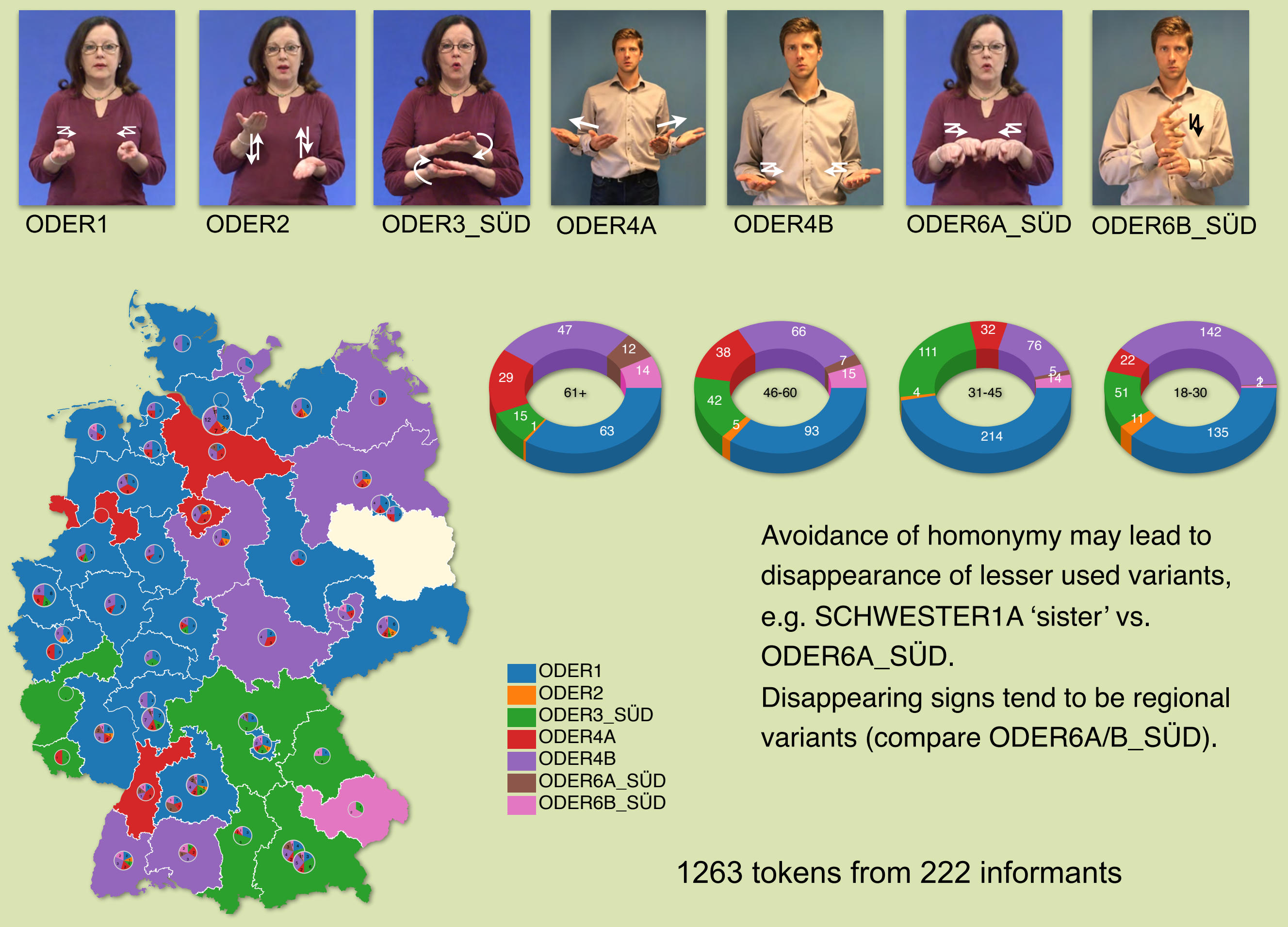
JETZT ’now’ age trend and hand shape



JETZT is a case showing younger signers’ preference for marked hand shapes.

1001 tokens from 229 informants

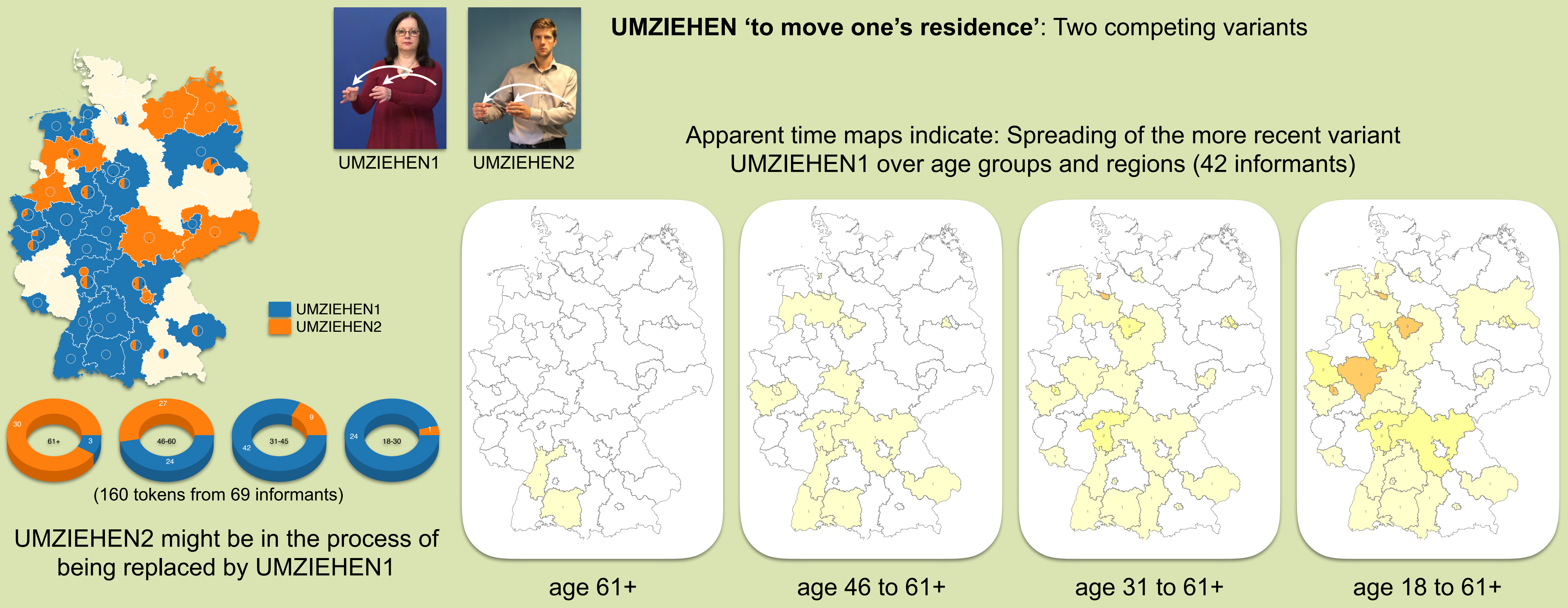
ODER ‘or’ regional and age distribution



Avoidance of homonymy may lead to disappearance of lesser used variants, e.g. SCHWESTER1A ‘sister’ vs. ODER6A_SÜD.
Disappearing signs tend to be regional variants (compare ODER6A/B_SÜD).

1263 tokens from 222 informants

UMZIEHEN ‘to move one’s residence’: Two competing variants



Apparent time maps indicate: Spreading of the more recent variant UMZIEHEN1 over age groups and regions (42 informants)

(160 tokens from 69 informants)

UMZIEHEN2 might be in the process of being replaced by UMZIEHEN1

age 61+ | age 46 to 61+ | age 31 to 61+ | age 18 to 61+

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies’ Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies’ Programme is coordinated by the Union of the German Academies of Sciences and Humanities.

References

- Bowie, David. 2005: Language Change over the Lifespan: A Test of the Apparent Time Construct. In: University of Pennsylvania Working Papers in Linguistics, Vol. 11, Issue 2, Article 3.
- Hanke,Thomas / Storz, Jakob. 2008: iLex - A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In: Crasborn, Onno / Efthimiou, Eleni / Hanke, Thomas / Thoutenhoofd, Ernst D. / Zwitserlood, Inge (eds.): LREC 2008 Workshop Proceedings. W 25: 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora. Paris: ELRA, 64-67.
- Nishio, Rie / Hong, Sung-Eun / König, Susanne / Konrad, Reiner / Langer, Gabriele / Hanke, Thomas / Rathmann, Christian. 2010: Elicitation methods in the DGS (German Sign Language) Corpus Project. Poster presented at the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, following the 2010 LREC Conference in Malta, May 22.-23., 2010. Workshop Proceedings. W13. 4th Workshop on Representation and Processing of Sign Languages: Corpora and Sign Language Technologies. May 22/23 2010. Valetta – Malta. Paris: ELRA. pp. 178-185.
- Gillian Sankoff. 2006: Age: Apparent time and real time. In: Elsevier Encyclopedia of Language and Linguistics, Second Edition, 2006. Article Number: LALI: 01479.

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