Variation of DGS lexical items

What sign language lexicography can gain from a mixed method approach: Corpus data supplemented by crowd sourcing

Gabriele Langer, Susanne König, Silke Matthes, Nele Groß, Thomas Hanke
Universität Hamburg, Institut für Deutsche Gebärdensprache und Kommunikation Gehörloser • www.dgs-korpus.de

DGS Corpus

Informants

- Controlled sample: balanced for region, gender, age-group
- Native and near-native signers, rooted in the Deaf community, regionally rooted (>10 years in the same region)
- No underage informants (year of birth: ≥1995) due to legal reasons
- Number of informants: 327 (filmed: 330)

Method

- Filmed conversations and staged communicative events
- Multi-modal corpus, lemmatised and accessible through iLex (Hanke/Storz 2008)

Data

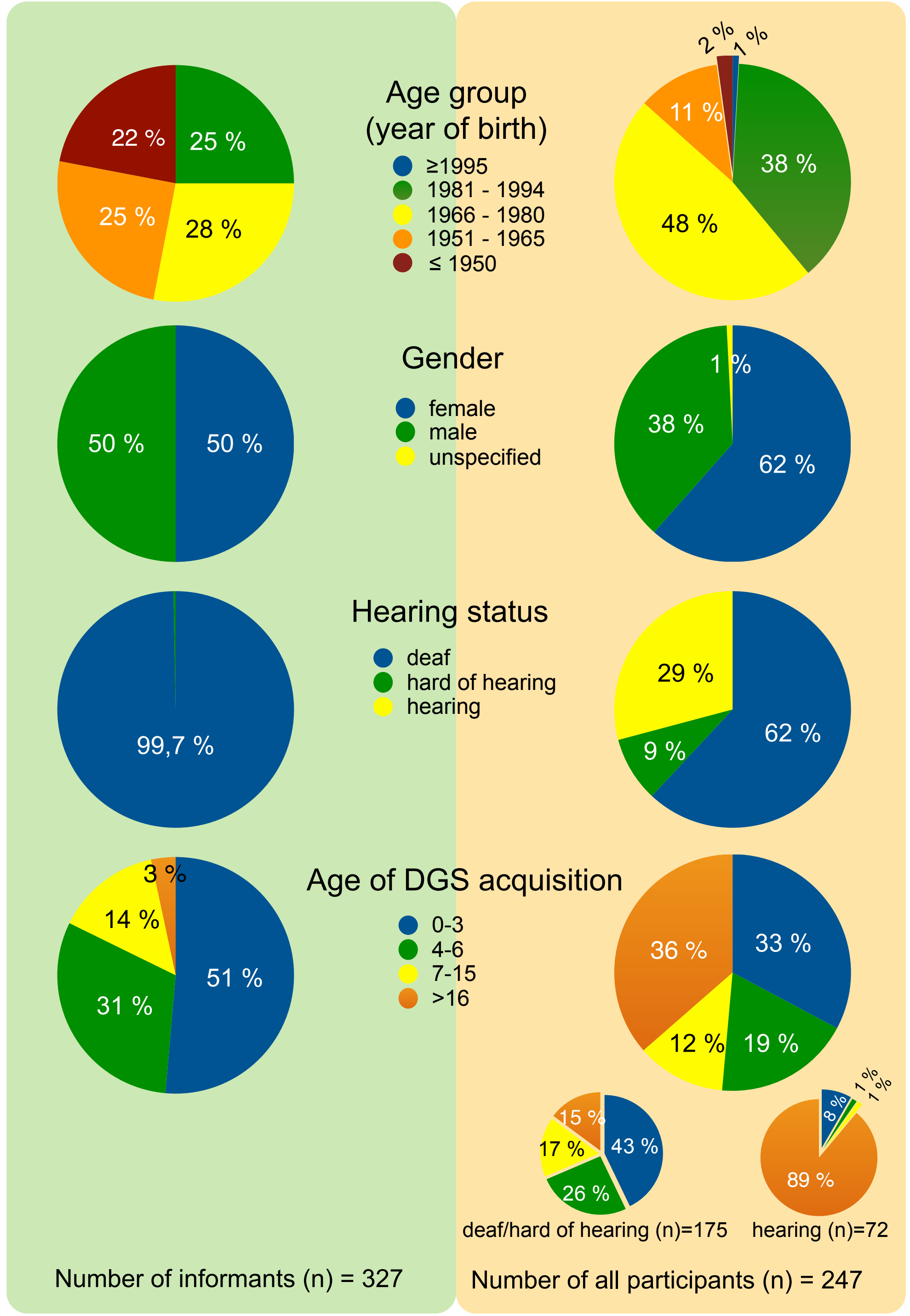
- Natural signing in context
- What is covered by data is largely uncontrolled and up to chance
- ≈ 616 hours of footage of relevant signing with an estimated 4 mio tokens
- Lemmatised (2015-12-09): ≈ 38 hours: 322,344 tokens

Uses for Lexicography

- Sign use in context: readings, usage, collocations, grammar
- Information on: frequency, typicality, distribution
- Authentic examples

Issues

- Data from few, but carefully selected signers
- Low frequency signs and sign uses may not be covered at all or not sufficiently for analysis (cf. Atkins/Rundell 2008, 57-61)
- Lack of data does not imply non-existence of item or phenomenon
- Time-consuming lemmatisation process
- Lemmatisation still in progress: regions not yet covered evenly
- Time-consuming annotation and analysis following lemmatisation



DGS-Feedback

Participants

- Uncontrolled sample, depending on volunteers and chance
- Deaf, hard of hearing, late-deafened, CI-users, hearing, all proficiency levels
- Biased toward people with an affinity for digital media: few elderly participants

Method

- Online survey in the DGS-Feedback survey system
- Questions presented both in DGS and written German
- Answers on isolated items (metalinguistic awareness)
- Items are voted either as ‘used’, ‘known’ or ‘unknown’
- Video comment option at all places in the survey
- Items are grouped in individual packages
- Results are immediately accessible via descriptive statistics and distributional maps

Data

- Controlled coverage: signs and readings can be presented as needed
- > 2000 items (≈ 1000 forms with ≈ 700 readings) online, 152,528 answers returned by 247 participants (2015-12-14)
- Positive (‘used’) and negative information (‘known’ and ‘unknown’), as well as information on passive vocabulary (‘known’), no missing data allowed

Uses for Lexicography

- Affirmation (or rejection) of items in question (e.g. forms, readings)
- Broader coverage of distribution (especially for region), supplementing corpus data
- Obtained information on candidates for citation form, variant forms, further readings

Issues

- Rationale: all members of the language community can participate and contribute
- Practical issue: recruitment of participants and ensuring continuous involvement proves difficult
- Online survey is still ongoing: constant preparation of new items and recruitment
- Way of presenting and the kind of stimulus may influence results

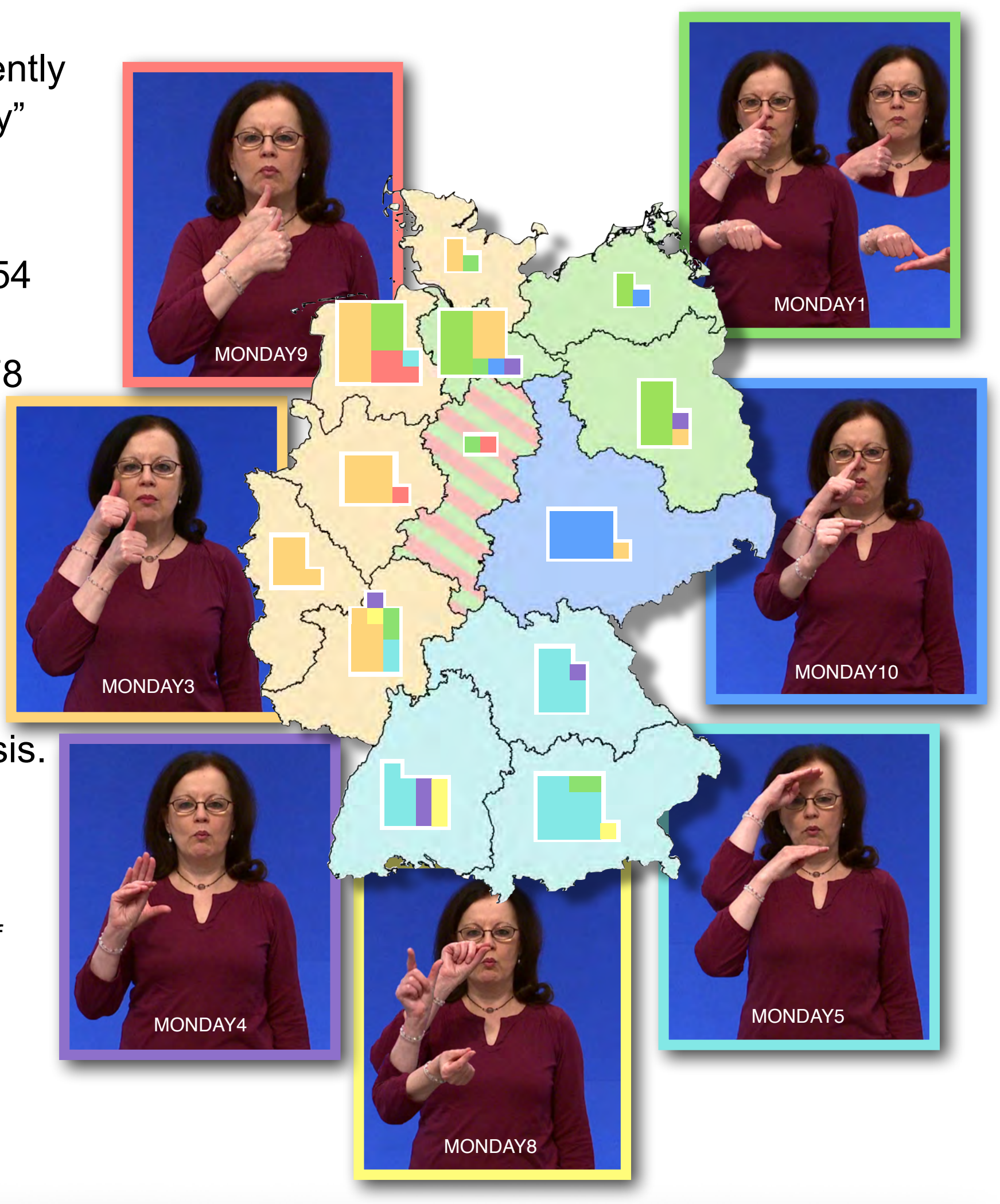
Combined method gain: Sociolinguistic variation – differences in region, age, gender?

Regional variation

The seven most frequently used signs for “monday” show a regional distribution.

144 corpus tokens by 54 informants and 108 used-answers of 78 feedback participants (together: data from 131 different people) have been analysed.

Some form variants have been merged into one sign for analysis. (2015-12-10)



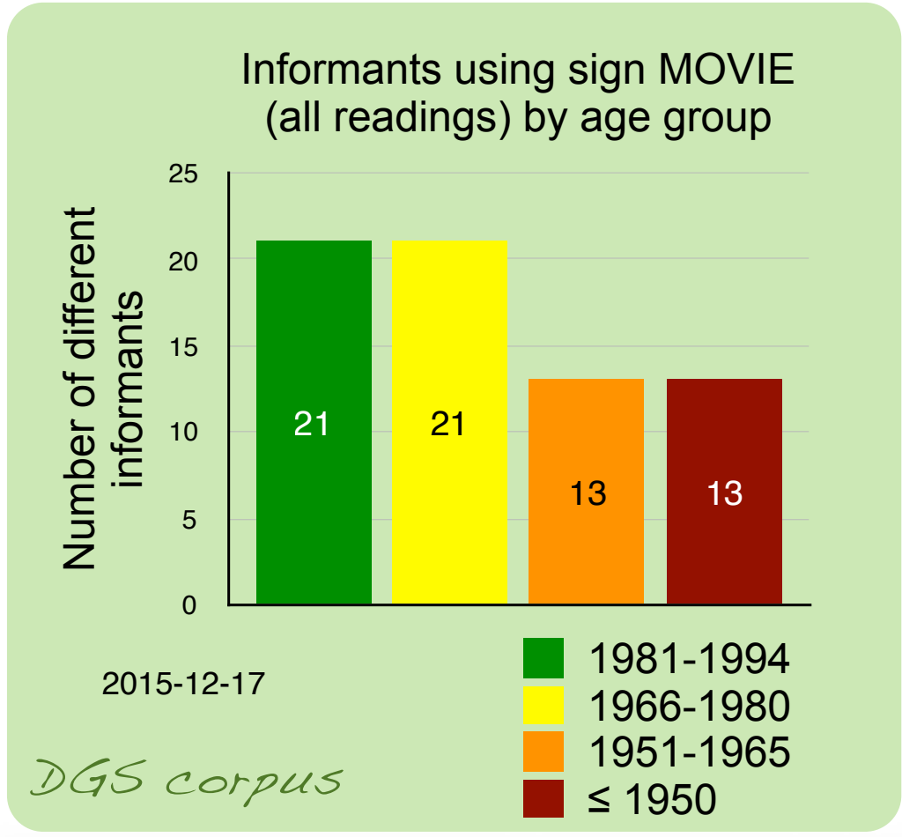
■ = evidence of regional use of a particular sign by 1 person (corpus tokens or feedback answer ‘used’)
total: 141 evidences

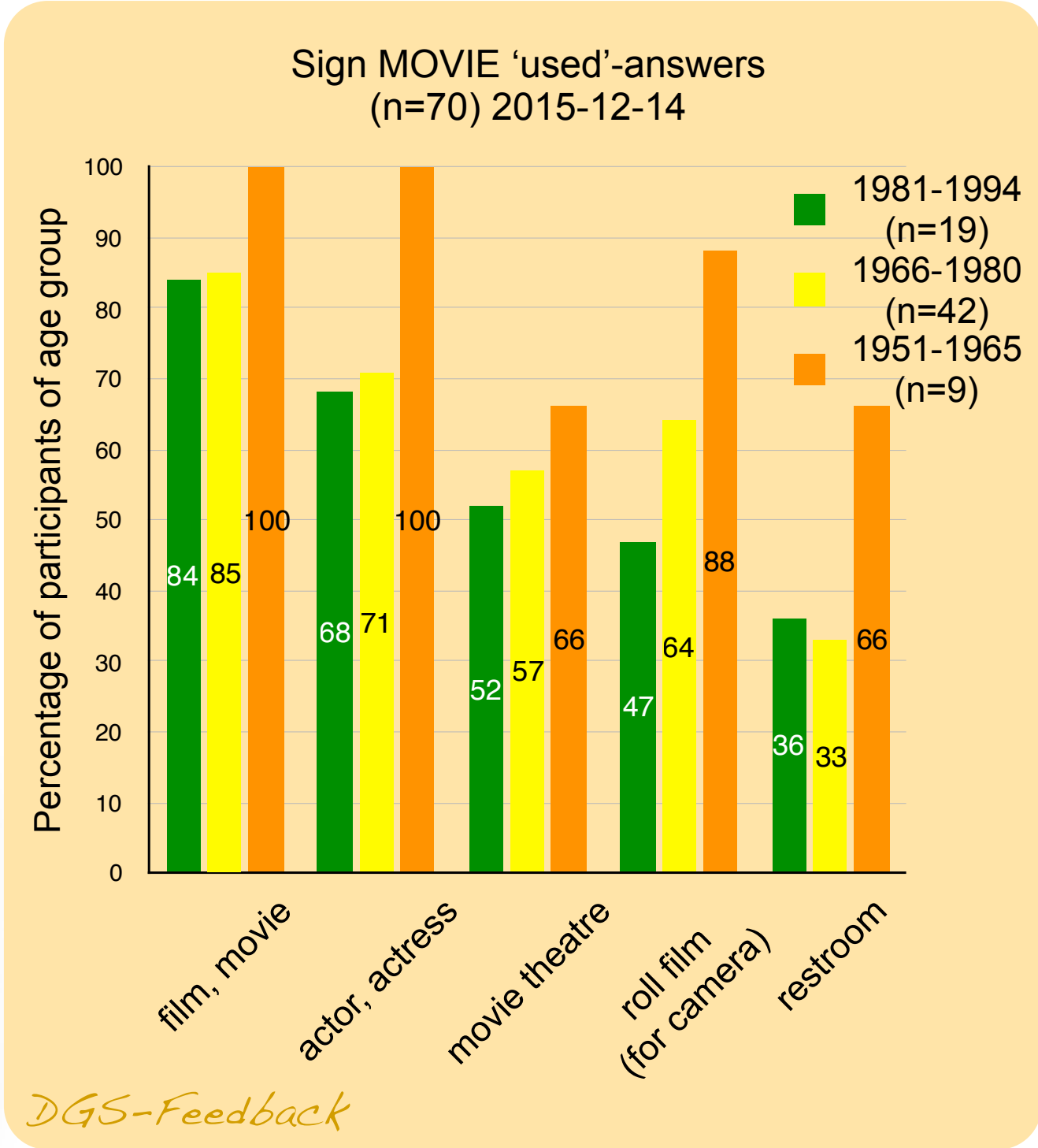
Possible dictionary information (example): The sign MONDAY5 is predominantly used in southern Germany.

Age variation

MOVIE

In the corpus more young than old people use the sign MOVIE. This may be due to an affinity of younger people to topics related to films and filming.
Feedback data, however, indicates that the percentage of younger people using the sign MOVIE in its various readings is smaller than that of older people. This may indicate language change in progress. Question: What signs are used by younger signers for these meanings instead?
The sign MOVIE will remain under close observation.



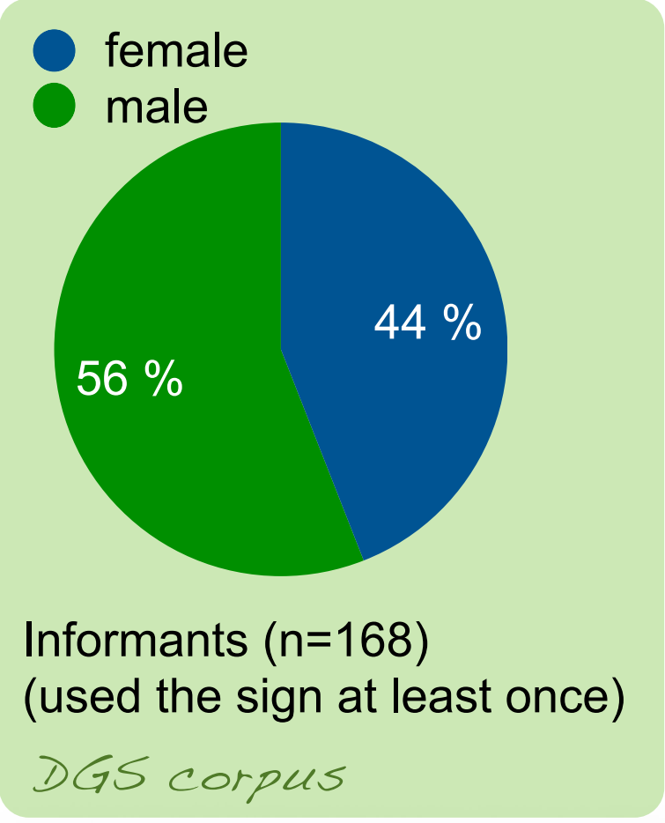


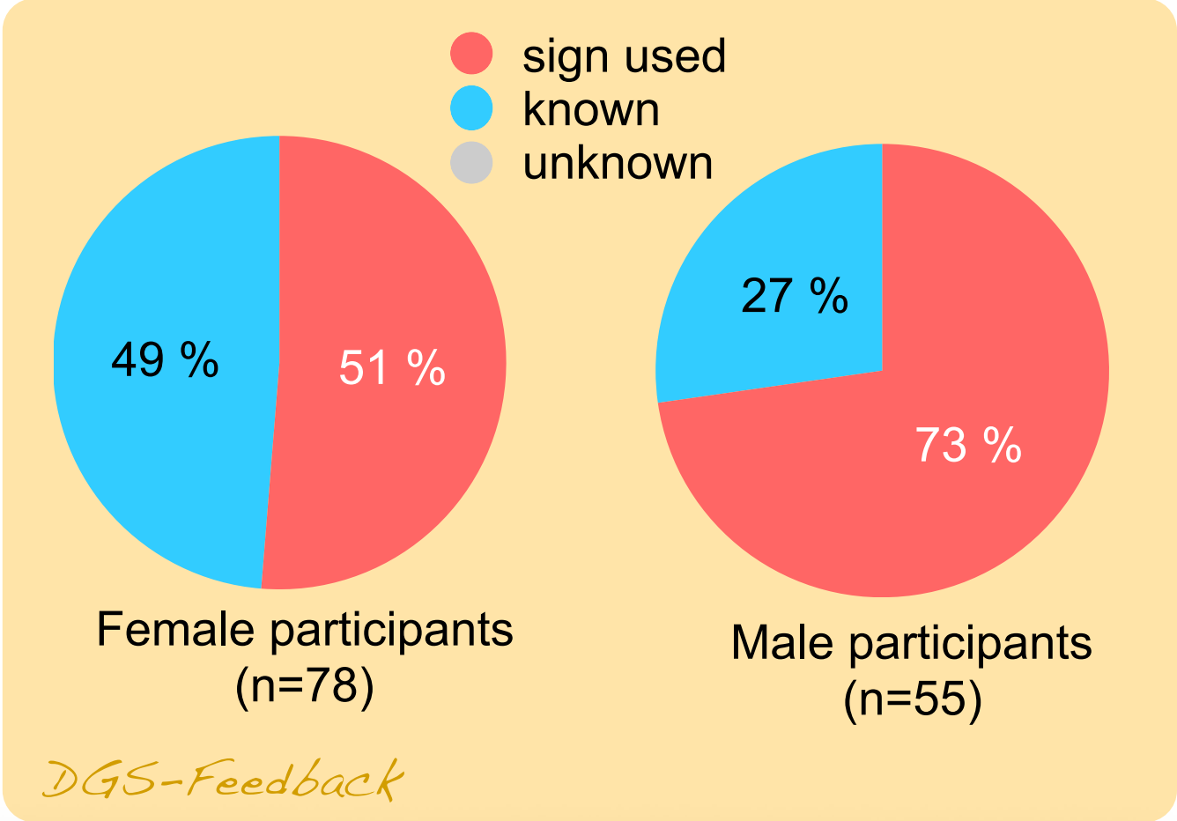
Possible information in a future dictionary: meanings “actor, actress”, “roll (film)”, “restroom”: dated.

Gender variation

WOMAN

This sign sketches the shape of a female breast. Some people consider this sign inappropriate for the meaning “woman”. Assumption: women, as those affected, use this sign less than men.
A comparison of data from both corpus (sign WOMAN) and DGS-feedback (readings ‘woman’ as in ‘female person’, ‘spouse’ and ‘Mrs.’) shows that women indeed tend to use this sign less than men, but that the sign is nevertheless widely used (441 corpus tokens of 168 informants). However, hearing participants in all answered ‘used’ for this sign much less (9 of 28: 32 %) than deaf/hard of hearing (72 of 106: 68 %) participants. (2015-12-17)

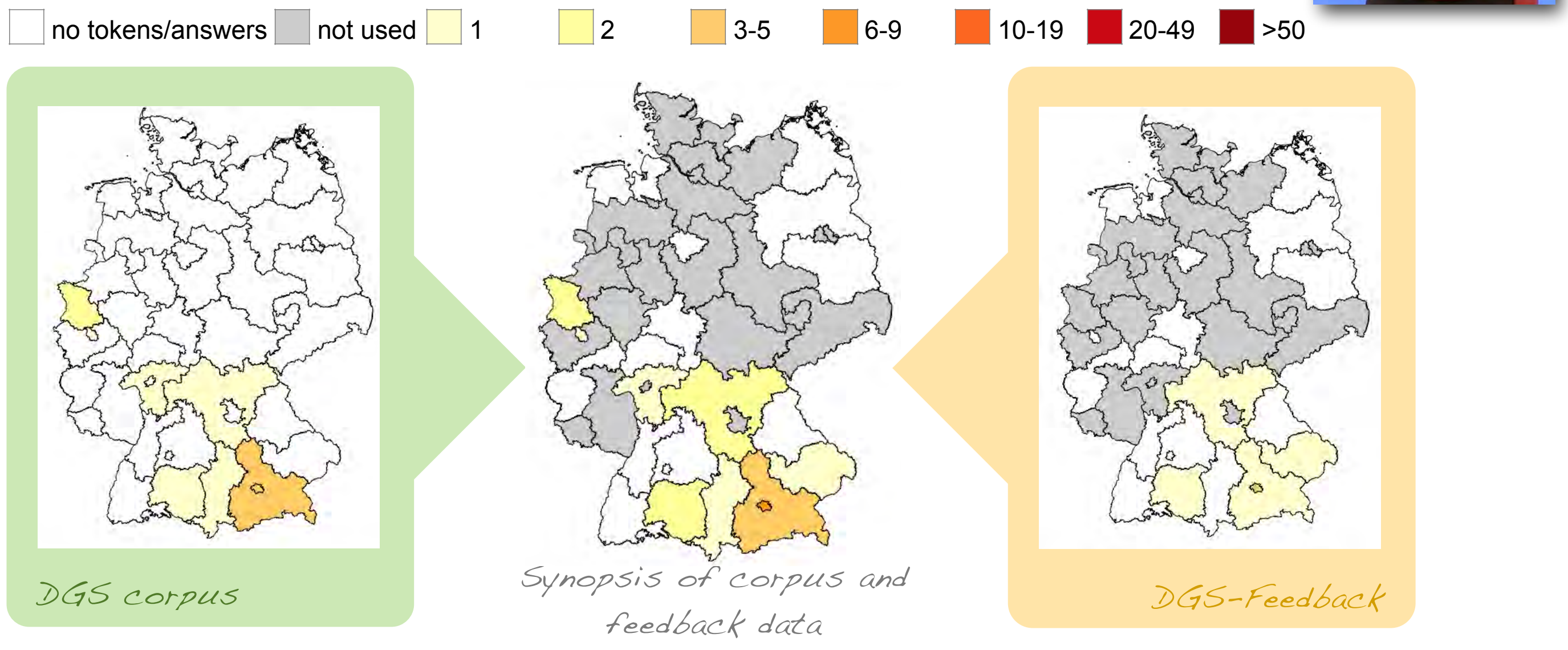




Possible dictionary information: the sign is widely used. Note: the sign may be considered inappropriate by some people, especially by woman.

Consolidation of regional distribution

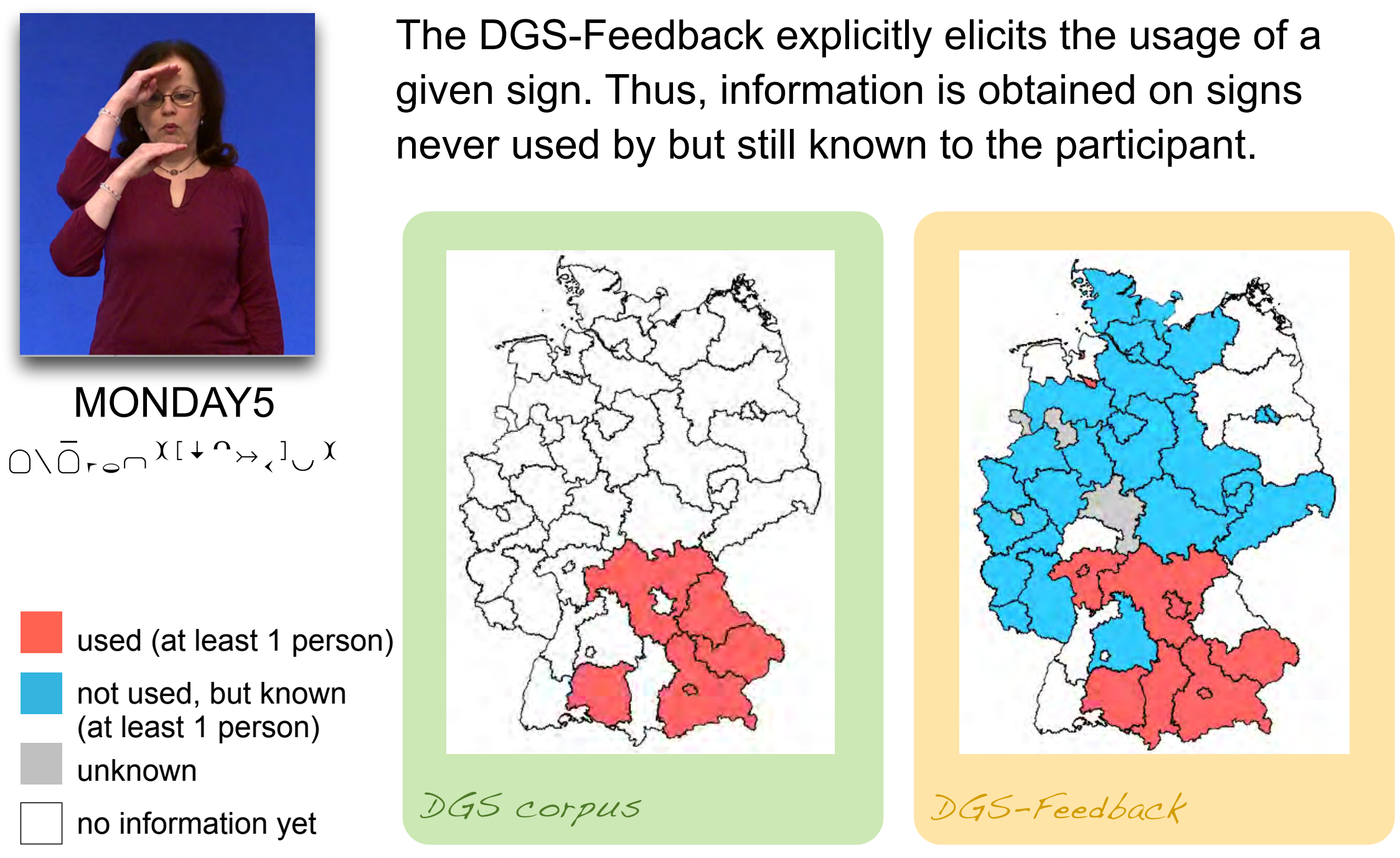
For the sign BLACK () corpus tokens show a regional tendency. This is confirmed and substantiated not only by the ‘used’-answers (yellow/orange range) from the DGS-Feedback but also by the negative answers (‘known’ and ‘unknown’: grey) from participants of other regions. Yellow/orange regions may also contain negative answers, but since the active use is in focus of the analysis, these are not displayed.



Active and passive vocabulary

MONDAY5

The DGS-Feedback explicitly elicits the usage of a given sign. Thus, information is obtained on signs never used by but still known to the participant.



Possible dictionary information: regional sign (southern Germany), but known widely across Germany.

Identification of meanings

MONEY

Readings (senses) with low corpus evidence can be tested via Feedback (here “cheap”).
Lexicographic aspect: Reading “cheap” has not been confirmed and is not a candidate to be listed in the dictionary entry.

Reading “money” in context with typical neighbours: GIVE, NEED, GET, EARN, LOT-OF, NONE, PAY, SAVE
DGS corpus

readings of MONEY	Corpus tokens	Feedback used answers (n=114)	Percentage Feedback used
money	342	107	94 %
financial	11	77	68 %
price (of goods)	19	63	55 %
expensive (price)	51	20	18 %
cheap (price)	1	1	1 %

Possible dictionary information: Readings: “money”, “financial”, “price”, “expensive”
Collocations (for “money”): GIVE, NEED, GET, EARN, LOT-OF, NONE, PAY, SAVE.

In written language lexicography analyses of large corpora as the basis for lexicographic descriptions are state of the art. Now sign language lexicography is on the brink of becoming corpus-based. This is an important and necessary step. But since corpus sizes of sign language corpora are considerably smaller than those of written languages, and especially while the DGS corpus is not yet fully lemmatised, it is very helpful to also use data acquired by a specifically devised online-survey to add to the picture of the distribution of signs, their variants and meanings. But also beyond the limitations of corpus size, we find a combination of methods productive and fruitful, since each method can answer questions on sign use the other cannot.
Corpus data is highly valuable for information on actual sign use in context, such as sign forms and variants, contextual meanings, collocations, grammatical behaviour and typical constructions. It can also be analysed for sociolinguistic and other factors of sign use (such as age, gender, region and register). The online survey can be used to add to and substantiate this information on distribution data, but in addition it can also provide other information on issues like passive language knowledge.
The above examples illustrate the advantage of combining two different methods of eliciting data from the language community and the resulting added valuable for lexicography.

References:

Atkins, B.T. Sue / Rundell, Michael (2008): *The Oxford Guide to Practical Lexicography*. Oxford, New York: Oxford University Press.

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.): *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*. Paris: ELRA, pp. 64-67. (URL: www.lrec-conf.org/proceedings/lrec2008/workshops/W25_Proceedings.pdf).

Kristoffersen, Jette H. / Troelsgård, Thomas (2012): „Integrating corpora and dictionaries: problems and perspectives, with particular respect to the treatment of sign language“. In: Workshop Proceedings. 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. Language Resources and Evaluation Conference (LREC) Istanbul, May 2012. ELRA, pp. 95-100. (URL: http://www.lrec-conf.org/proceedings/lrec2012/workshops/24.Proceedings_SignLanguage.pdf)

Langer, Gabriele (2012): ”A Colorful First Glance at Data on Regional Variation Extracted from the DGS-Corpus: With a Focus on Procedures“. In: *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon Language Resources and Evaluation Conference (LREC) Istanbul, May 2012*, pp.101-108. (URL: http://www.lrec-conf.org/proceedings/lrec2012/index.html)

Langer, Gabriele / König, Susanne / Matthes, Silke (2014): ”Compiling a Basic Vocabulary for German Sign Language (DGS) – lexicographic issues with a focus on word senses“. In: Abel, Andrea / Vettori, Chiara / Ralli, Natascia (eds.): *Proceedings of the XVI EURALEX International Congress: The User in Focus, July 15-19 2014 in Bolzano/Bozen* – Italy, pp.767-786. (URL: http://euralex2014.eurac.edu/en/callforpapers/Documents/EURALEX%202014_gesamt.pdf)

Matthes, Silke / Langer, Gabriele / Blanck, Dolly / Hanke, Thomas / Konrad, Reiner / König, Susanne / Regen, Anja (2013): „Involving the crowd: How to complement corpus data in the process of dictionary making“. Abstract submitted for TISLR 11, July 10-13, 2013, London. (URL: http://www.ucl.ac.uk/dcal/tislr/abstracts/tislr11_submission_239.pdf)

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“. In: *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. (URL: http://www.lrec-conf.org/proceedings/lrec2010/index.html)

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.

Poster presented at the 12th International Conference on Theoretical Issues in Sign Language Research (TISLR12), Jan 4-7, 2016 at Melbourne, Australia. (Corrected Version 2016-01-19).