A Colorful First Glance at Data on Regional Variation Extracted from the DGS-Corpus: With a Focus on Procedures

Gabriele Langer

University of Hamburg
Institute of German Sign Language and Communication of the Deaf
Binderstr. 34, 20146 Hamburg, Germany
E-mail: gabriele.langer@sign-lang.uni-hamburg.de

Abstract

In this work in progress procedures for analyzing and displaying distributional patterns of sign variants have been developed and tested on data for color signs elicited by the DGS Corpus Project. The data for this preliminary study were elicited as isolated signs and have been made accessible through spot annotations in iLex. The annotations had not been lemma revised but nevertheless revealed some interesting insights. Several color signs exhibited a high degree of variation. The distributional maps showed that a number of signs were mainly used in certain regions and thus provided evidence on dialectal differences within DGS. The relevant information necessary to generate distributional maps have been directly extracted via SQL-statements from the corpus and fed into R. The approach is data driven. The distributional maps show either the distribution of one sign form (variant) or of several different variants in relation to each other. Analyses of regional distribution as displayed by the distributional maps may support the annotation and lemma revision process and are a valuable basis for a lexicographical description of signs and their use as needed for compiling dictionary entries. A refined procedure to take multiple regional influences on informants into account for analysis is proposed.

Keywords: generation of distributional maps from corpus data, regional variation in DGS, color signs, data-driven approach

1. Introduction

Within the DGS Corpus Project about 1160 hours of footage with an estimated 540 hours of signed activity have been collected. 330 informants in 13 German regions were filmed in pairs. This material will constitute a general corpus of German Sign Language (DGS) after it has been made accessible through annotation. The next stage of the project is dedicated to annotation and transcription of the raw data. At a later stage the first corpus-based general dictionary of DGS–German will be produced based on the data documented in the corpus.

One of the project's aims is to document lexical variation including regional variation. Information on regional variation is an interesting and useful piece of information on signs that should be included in dictionary entries wherever possible. Within the project, procedures need to be proposed, tested and established to extract and present information on regional distribution from the corpus data efficiently as it is needed to support the compilation of dictionary entries. Even though the prerequisite for the analysis of many sociolinguistic variables are provided for in the metadata gathered, these kinds of general studies on variation are not part of the DGS Corpus Project itself. Within the project, only variation of individual signs is analyzed as far as this information is needed for the compilation of a dictionary entry such as the sign's regional distribution or sign use restricted to certain age groups.

Since annotation is currently in progress, analyses on regional distribution of signs from the corpus cannot be based on large amounts of empirical data yet and therefore can only be preliminary. To gain practical experience in dealing with widespread variation spot annotations of color signs filmed during the task *elicitation of isolated signs* are being used as a testing ground for analysis procedures.

2. Elicitation Method

One of the two elicitation tasks specifically aimed at eliciting regional variation is the *elicitation of isolated signs* (cf. Nishio et al., 2010). The goal was to elicit signs for a small number of selected concepts from a large number of informants. In this task concepts that were known to exhibit a high variation in DGS were presented as written words, some of them also in combination with a picture. Informants were asked to produce their signs for these concepts. Eleven colors (red, blue, yellow, green, orange, purple, pink, brown, black, white, gray) were presented on the screen as unicolor plane without written references to the concepts. Informants were asked to name these colors.

3. Sample Size

One informant of each pair (i.e. 165 informants) was asked for his/her color signs in the task *elicitation of isolated signs*. For preliminary analysis raw data from 156 informants of 12 regions available were transcribed resulting in 2052 tokens for colors. This included the tokens from the spot transcription[1] of the *isolated signs task* and tokens that have already been annotated within other parts of the corpus material. The movies from the

[1] Spot transcriptions for this study were made by Nele Groß, Ilona Hofmann, Lutz König and Gabriele Langer. Technical support was provided by Sven Wagner.

last region (Leipzig) and a few movies from other regions had not been available for transcription at the time and could therefore not be included. Even though the sample size is rather large it is still too small to gain a clear picture of regional distribution for all variants, especially since other factors like schooling might have a greater influence on variant use than the actual place of living. However, the preliminary results show some interesting tendencies of regional distribution. Within the DGS Corpus project a web-based feedback function (technical term: voting) is planned and in the future will provide further information to be included in the analyses of regional distribution of signs.

4. Annotation

The data of this study have been annotated in a very basic way with the transcription tool and integrated database of iLex (Hanke, 2002; Hanke & Storz 2008). Spot annotations have been carried out to identify different form variants for color signs. All variants have been described by separate type entries regardless of whether they would be considered phonological or lexical variants. Forms e.g. with a clearly extended thumb constituted new type entries in iLex whereas small deviations of form that have been known to occur frequently with certain handshapes (such as small differences of thumb position or more or less spreading or bending of fingers) or that seemed to be either idiosyncratic or accidental did not constitute new type entries. Instead these minor differences were noted with the token (i.e. in the token tag) as form deviations from the citation form of the type. When the number of tokens with the same deviation within a type entry is increasing they can be re-categorized at a later stage of the annotation process called lemma revision (cf. Konrad, 2011 pp. 93-96; König et al. 2010). Also, some kinds of variation that have led to separate entries in one case (such as one-handed vs. two-handed) have been subsumed under one entry in other cases with qualifications or token deviations noted.[2] This is to say, the data is still somewhat messy as it has yet to undergo the lemma revision process. Therefore the categorizations of this study are preliminary. It is expected that some form types will be merged into one while others (for example BLACK1) may be separated in two or more types on grounds of the distributional data of form variation so far considered as minor. For this preliminary analysis of regional distribution all variants have been annotated and analyzed separately focusing on the variants with the highest number of tokens (9 tokens or more) and leaving out variants with a lower number of tokens. The point of this preliminary study is to show that even with corpus data that is not completely consistent yet analyses of distribution can provide some useful insights that may even support the decision-making process of re-categorizing the data.

[2] In the DGS Corpus Project the iLex database and working environment is used for annotations. The database contains large amounts of annotated data and type entries from previous projects. Each project had used somewhat different annotation rules. Annotation guidelines, structures and procedures for the DGS Corpus Project are still being developed. To draw on type entries from previous projects is a huge advantage but also constitutes a challenge for the consistency of rule application. While the number of hands had often constituted new type entries in the past the number of hands are now being annotated by qualifier structures implemented in iLex (see Konrad et al. 2012, this issue). This is the reason why for some color signs there still exist separate entries for one-handed and two-handed variants while for others this kind of variation is already marked by qualifiers within the same type. Re-categorizing old entries and tokens following new annotation rules and structures will take some time and effort and will happen step by step as new rules are being developed and implemented and more and more sign entries go through the lemma revision process.

5. Analysis of Distribution

5.1 Regionality of Informants

One requirement for the selection of informants was their rootedness within a given region. Only lifelong or at least long-term residents of a region were accepted as informants. Preferably the informants should have grown up and currently have their permanent residence within the region. A residency of at least ten years within the region was also accepted. Metadata of the informants include the place of living, the place of growing up, the school they attended and all other places the informants had been living at for a longer period of time.

Three informants who had recently moved away were nevertheless accepted for their original region. In this case the current place of living did not coincide with the prominent regional linguistic affiliation of the informant. For these informants their last residence within the original region has been used for the preliminary analysis of regional distribution.

5.2 Displaying Regional Distribution

This first preliminary study is based on the place of residence of the informants. The distribution of the most frequent color signs (9 tokens or more) was matched onto the map of Germany with a resolution at the county level. For this each informant's place of living was matched to the corresponding county and the county coding (corresponding to the GADM dataset for Germany[3]) was stored as metadata to the informant within iLex. By an SQL query all county codes with an attested sign use for a certain sign were extracted from iLex. All counties with attested sign use were then colored to show the regional distribution of the sign in question. The data exported from iLex were fed into the statistical analysis program *R* using the packages *maps* and *sp* and the GADM dataset for Germany to produce the maps.

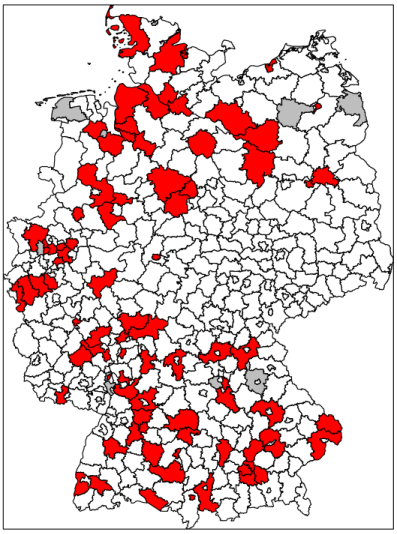
The maps displaying the attested use of a specific sign are a result of the described procedure and directly driven by the data from the corpus, combining metadata (place of living) and annotation data. Maps can either

[3] http://www.gadm.org/

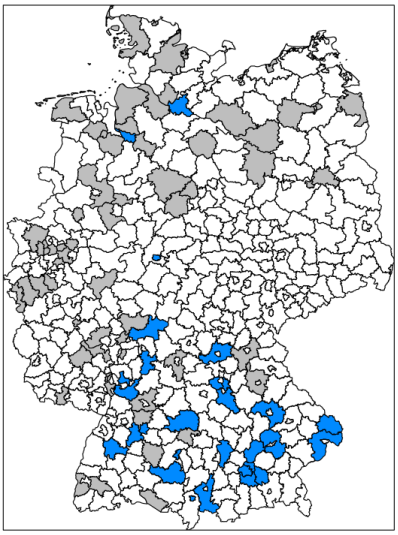
show the distribution of one particular sign as in map 1 (RED1) and map 2 (BLUE3) or the distribution of two (map 3) or a number of variants in relation to each other (maps 4 and 5).

In the future the described procedure can be implemented to automatically produce distributional maps of selected signs on command and thus provide a quick overview to support lemma revision and the compilation of dictionary entries.

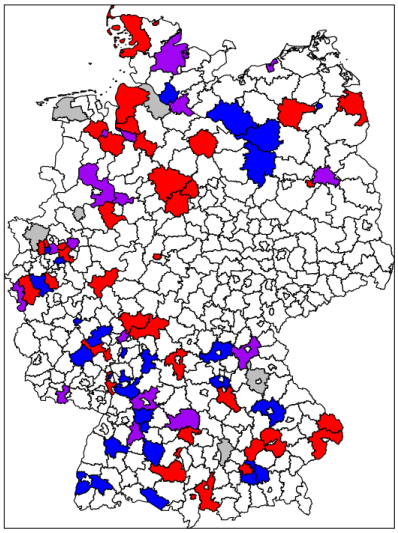
5.3 Distributional Maps



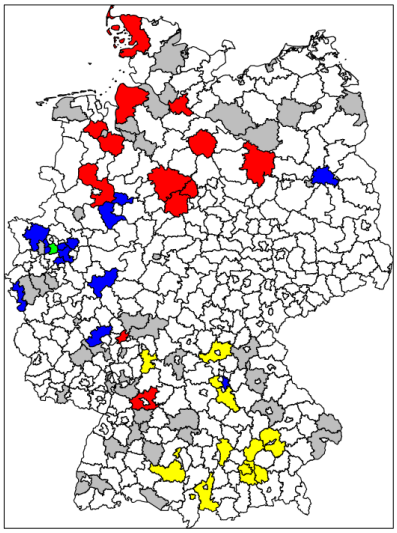
Map 1:
RED1



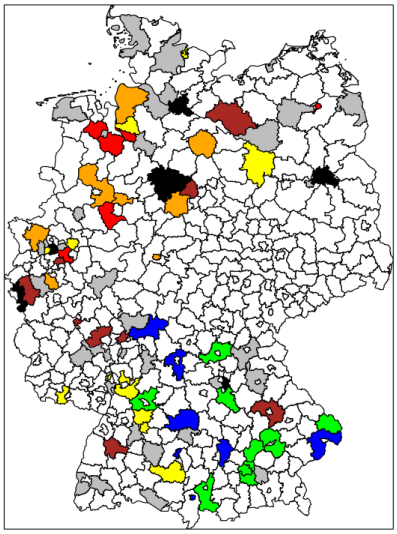
Map 2:
BLUE3



Map 3:
BLACK1 (red),
BLACK2 (blue)



Map 4:
GREEN2 (blue),
GREEN3 (red),
GREEN9A (yellow)



Map 5:
six variants for *brown*
(glosses and HamNoSys for these variants are listed on the right)

The maps are based on data of 156 informants from 90 counties. Counties without informants are colored white, counties with informants but no attested sign use are colored grey.
Maps 1 and 2 show the distribution of a single variant. Map 3 contrasts the use of two variants while map 4 contrasts the use of three variants. In both maps overlapping areas of use are marked by the corresponding mixed color, e.g. areas of overlapping use for BLACK1 (red) and BLACK2 (blue) are colored purple (map 3).

Map 5 shows the distribution of the following six variants for the color *brown*:
BROWN2A (red),
BROWN029 (orange),
BROWN7 (yellow),
BROWN8 (brown),
BROWN9 (blue),
BROWN4 (green),
(overlapping areas of use: black)

5.4 Results

For this study 2052 tokens of color signs from 156 different informants of 90 counties in 12 regions were annotated and matched to 256 types. For 117 of these types only one token was found, 45 types had 9 tokens or more. Only these more frequently used types were analyzed for regional distribution. They accounted for 75%

of all tokens. (For an overview on the numbers of types and tokens see table 1).

mouthing: purple*: lila; purple**: violett; pink*: rosa; pink**: pink	number of types (variants)	number of tokens	types with one token	types with 9 or more tokens (A)	number of tokens of (A)	For (A): % of all tokens
blue	23	173	8	6	138	80%
brown	34	161	16	7	93	58%
yellow	32	192	18	5	152	79%
grey	47	169	19	5	72	43%
green	39	182	19	4	98	54%
purple*	23	174	5	3	126	72%
purple**	2	7	0	0		
orange	21	177	12	5	153	86%
pink*	26	160	10	3	107	67%
pink**	6	11	3	0	0	0%
red	4	163	1	1	154	94%
black	4	310	0	2	298	96%
white	13	167	5	4	148	89%
beige	1	5	0	0		
turquoise	1	1	1	0		
	276	2052	117	45	1539	75%

Table 1: Number of types and tokens for colors

Results of this preliminary study show that there is a lot of variation in color signs in DGS. Even though the data still has to undergo the lemma revision process it nevertheless can already be used to visualize tendencies of distribution. Five examples of distributional maps for selected color signs are included in this paper. The maps show that RED1 (map 1) is used all over Germany (as far as data was available for these areas) while BLUE3 (map 2) is primarily used in Southern Germany. BLACK1 and BLACK2 (map 3) both seem to be used in all areas of Germany. The overlap areas of attested use are marked by the corresponding mixed color (in this case purple as the mixture of red for BLACK1 and blue for BLACK2). A closer investigation of the form deviations of BLACK1 may bear interesting results as a variant with slightly spread and bent fingers appears to be used in Southern parts of Germany. Map 4 is an example of a very clear regional distribution of three lexical variants for *green* (GREEN2, GREEN3 and GREEN9A). Map 5 shows the distribution of 6 variants for *brown*. Here overlap areas are colored black. Maps 3, 4 and 5 all indicate that there might be a distinct dialectal area in Southern Germany while dialectal areas in other parts of Germany cannot be seen as clearly from these few analyses. It will be very interesting to look at signs from other domains and also from the data elicitation region of Leipzig to get a clearer picture of dialectal regions of DGS in Germany.

5.5 Limitations of the preliminary study

This preliminary study has a number of limitations. The analyzed sample does not include data from all regions and informants yet. The informants filmed at Leipzig (from an area covering the Southern part of former East Germany) are not included. Also in other data collection tasks further tokens of color signs will occur that have not been transcribed yet. More data is needed to stabilize the findings and to fill the gaps.

All annotations for this preliminary study have to undergo lemma revision. Within this review process some variants will probably be divided into different subvariants. For example, the deviation information of the tokens of BLACK1 indicate that there may be at least one subvariant that is consistently used in the south. Other forms (especially forms with only one or few tokens) might be re-categorized as deviations of other variants thus reducing the number of variants for the associated color. This is to say that the results presented in this paper indicate tendencies but are to be received with caution and not to be taken as final results.

The chosen geographical display of regional distribution has also some limitations. Berlin has been treated as one area (county), but for historical reasons should be divided into an Eastern and Western part to be able to analyze effects of the division of Berlin from the 1960s to the 1980s on sign distribution in that area. Some recent changes of administrative areas (counties) are not included in the GADM dataset and one county is completely missing. For future implementation of this procedure a more complete and up to date dataset has to be used.

The number of tokens or the number of different informants per sign and county respectively are not displayed on the distributional maps yet. Including this information would show the central areas of use more clearly. Improved versions of distributional maps should also indicate overlap areas more clearly.

Other regional influences than the place of living should be taken into account. See section 6.3 for a suggested approach to this issue.

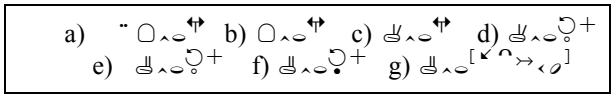
Sociolinguistic variables other than region should also be investigated and put into relation to regional factors as it was done in other projects on sociolinguistic variation in signed languages (cf. for example Lucas et al. 2001; McKee & McKee, 2011; McKee et al., 2008; Schembri et al., 2009). As this is not part of the DGS Corpus Project, this issue awaits further research.

6. Issues of Procedure and Research

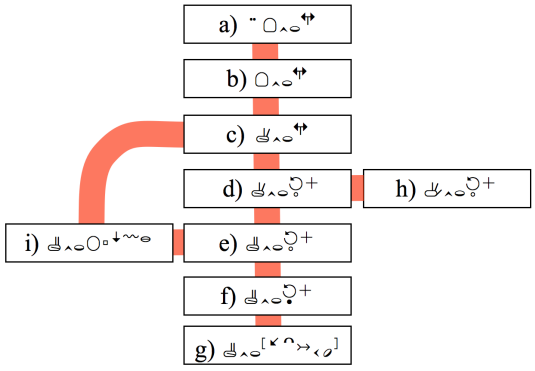
6.1 Lexical and phonological variation

In the annotation and analysis of variants, usually lexical variation (phonologically unrelated forms, that is, distinct signs) is distinguished from phonological variation (phonologically related forms of the same basic sign). Two similar sign forms are generally treated as phonological variants (also called subvariants) rather than

lexical variants when they differ in only one parameter from each other, such as handshape or movement or place of articulation (cf. for example Lucas et al., 2001 p. 180; Johnston, 2003 p. 349; McKee & McKee, 2011 p. 502; Hollman & Sutrop 2010 p. 141). However, the distinction between phonological and lexical variation is not always as clear as it might seem on first glance. Sometimes there exist chains of sign forms where each sign form differs from its neighbors only slightly in one formational feature, so that direct neighbor(s) in the chain would usually be considered phonological variants of each other, while the signs at the distant ends of such chains may not have much in common with each other and would usually not be analyzed as phonological but rather as lexical variants of each other (see example 1 for a chain of partly similar forms used for *blue* in the DGS Corpus data).



Example 1:
A chain of partly similar forms used for *blue*



Example 2:
Partly similar forms used for *blue* (branching chain)

Examples 1a and 1g seem to be totally unrelated sign forms and differ with respect to number of hands, handshape and movement. However, in between these two signs other forms exist where each sign in the chain differs from its neighboring signs with respect to only one formational feature: a to b: number of hands, b to c: handshape, c to d: movement, d to e: handshape, e to f: size of movement, f to g: shape of movement (arc instead of full circle with an additional change of orientation making the arc anatomically more comfortable). Even if for this reason 1g would be considered distinct from the other forms, the same point could be made focusing on a and f. To complicate things further, chains may also branch off and possibly reconnect (see example 2).
This example shows that distinguishing phonological from lexical variants cannot be based on the formational similarity of the sign forms alone. König et al. (2008, p. 394) suggest to take into account the underlying image and the image producing technique of signs when determining whether two similar forms are phonological variants of the same sign (based on the same underlying image, produced by the same technique) or independent lexical variants (different underlying images and/or techniques). This can be helpful when dealing with iconic signs, but it cannot be applied when the signs in question either lack iconicity or when their underlying image cannot be determined, as it is the case for many color signs in DGS.
In the case of this study one-handed and symmetrical two-handed productions were often treated as the same sign (example 1a and b), as well as certain differences in the spreading of fingers that often occur in signs with specific handshapes (example 1d and 1e, also flat and slightly spread fingers for the B-handshape in BLACK1). Frequency of occurrence can be taken as an additional criterion for grouping tokens into separate entries. Frequently attested forms were treated as separate entries while others that had only one or a few tokens used by only one or few signers were either interpreted as idiosyncratic deviations of another form (for example 1f was interpreted as instantiations of 1e with the deviation of an enlarged movement) or they have been omitted in the overall analysis because their number of tokens was too small.
The analysis of regional distribution of very similar forms may reveal whether they are different phonological variants of the same sign used in the same region or two dialectal variants used in different regions. Thus data-driven distributional maps as introduced in this paper may aid the annotation process itself by providing clues for categorizing or re-categorizing certain form variants into one or separate entries of the lexical database used as a basis for annotations. For the lexicographical description of individual signs these analyses are also very helpful. Phonological variants with the same distribution might better be treated in one common sign entry in the dictionary covering these forms and describing the range of the variation while it would be more user-friendly to produce two separate sign entries for dialectal variants. Distributional maps can also support practical lexicographic work for identifying and describing the use of individual signs and some smoothed-out version of the maps could even be included as a visual hint on the distribution of the given sign.

6.2 Multiple Regional Influences

Depending on where DGS was acquired the place of growing up or living might not be the strongest or the only regional influence on the signing of a particular informant. For example, it was reported for many sign languages that residential schools have a strong influence on the signs a signer uses (cf. for example Lucas et al., 2001; Schembri at al., 2009; Schermer, 2003). Studies on regional variation of spoken languages usually only include informants who have lived all of their lives in one place/area. For sign languages it is rather unlikely that a sufficient number of such signers can be found and

recruited. Therefore also signers with a long but not a livelong residence in the specific area are accepted as informants, even though their signing may show influences of different regions.

When several geographical data have been collected on each informant it could be attempted – provided the sample size is large enough – to take different geographical influences on a particular informant into account for an analysis of the distribution of a certain sign. This could be done by comparing the different geographical regions attributed to the informant to the overall regional distribution of the given sign form by other informants and identifying the most plausible regional influence for the given signer and sign. In the next section (6.3) a procedure for this kind of analysis is outlined. This type of analysis will become especially useful when dealing not only with corpus data but also with data collected through the public feedback gathered at a later stage of the project.

The public web-based feedback function will supplement the data from the corpus. Within this feedback function members of the sign language community are asked to participate and answer questions on the signs presented there. The feedback will include information such as whether the participant knows and/or uses a particular sign or not. One has to register in order to participate. Registration will include some geographical information about the participants such as place/region of living and possibly other geographical information like place/region of schooling or place/region of growing up. It is expected that a number of participants have been living in several different regions and that each of these may have influenced their signing and their knowledge of signs.

6.3 Dealing with multiple regional influences: proposed procedure

Here an analytic procedure is outlined of how to take multiple potential regional influences on one informant into account for regional analysis of a particular sign. This outline is meant as a contribution open for discussion as it is work in progress and has not yet been implemented or tried out. The idea is that a particular informant may have several regions that potentially influence his or her signing, for example region of growing up, region of schooling, region where his or her deaf parents come from, different regions of long-term residence, long-term stay abroad and so on. In this paper these regions are called potential regional influences (PRI). All PRIs of an informant have to be known and matched to a geographical area. They also have to be categorized for their kind (e.g. permanent residence, place of schooling, place of growing up and so on). Provided enough data is available from many other informants using the same sign it should be possible to identify the most probable regional influence (MPRI) of the given PRIs for the use of this particular sign by comparing the PRIs to the attested regional distribution of the sign.

The analysis procedure can be described as follows:

Step 1: As basis for the comparison all areas of interest (for example all counties of Germany[4]) are given a value for the sign in question – depending on how many tokens of the sign from how many different informants are attested and attributed to this area. I will call this set of values for each area a-values. All PRIs of all informants are to be taken into account for this a-value calculation for a particular sign. When one informant has three PRIs attributed to him/her and uses a certain sign, then this contributes to the a-value of all three PRIs (e.g. counties). Areas with many tokens from many different informants receive a high a-value (e.g. 4), areas with few tokens from only few different informants receive a middle a-value (e.g. 3), areas with tokens by only one informant receive an a-value of 2 and areas that have no tokens but are neighboring a high or middle score area receive a low a-value (e.g. 1).[5] All other areas receive the a-value of 0. All areas with an a-value above 1 are called attested areas, all areas with the value 1 are called neighboring areas. Threshold values need to be defined for this categorization as high or middle score attested area. The threshold values can be adapted to the number of overall tokens of the sign.

Step 2: The a-values are taken as basis to determine the most probable PRIs for all informants and their tokens. Now all PRIs of each informant in question are compared to the a-values of the areas and the most probable area of influence for this sign may be determined by the following rules:

- a) The PRI area that has the highest corresponding a-value is the most probable influence for the use of the sign in question.
- b) When two or more PRI areas have the same corresponding a-value, the PRI area with the highest priority on a priority list (see below) is chosen as the most probable.
- c) When no PRI area has a corresponding a-value above 2, then the PRI area with the highest priority on a priority list (see below) is chosen as the most probable.

In order to resolve cases where two or more PRIs have the same value (see above case b and c) a priority list has to be defined that ranks the kinds of geographical areas (for example: area of growing up is favored over area of only two years of residence). This list ensures that for each sign and informant exactly one area of the PRIs can be chosen as the most probable even if there are only few tokens available or if none of the PRIs of the particular informant overlaps with the PRIs of other informants.

Once the most probable area (MPRI) has been determined for a given sign and informant of his or her PRI areas, all tokens of this sign by this informant are attributed to the determined MPRI.

[4] As we do not have data from all counties it might prove more useful to broaden the granularity from counties to larger areas such as districts. In this case the procedure can be adapted accordingly.

[5] In addition, PRIs of informants with a livelong residence at one place and therefore only one PRI should rank higher than the PRIs of informants with several PRIs.

Step 3: The values of all areas (e.g. counties) are again determined. This is done on the basis of all identified most probable areas (MPRIs) only. This new set of values for all areas will be called b-values.
Step 4: The results of step 3 can be displayed on a map using different shades of colors for high, middle and low b-value areas.
The described procedure will consolidate the areas of attested sign use and filter out most accidental singular occurrences. Another advantage of this procedure is that competing signs for the same concept used by the same informant can be taken into account and analyzed separately. Other studies have used only the first response of an informant to a lexical elicitation task for analysis because it was considered "the signer's default, spontaneous usage" (McKee & McKee, 2011 p. 499). However, it is likely that within a corpus of spontaneous signing one informant uses several competing variants without one variant being more spontaneous than the other. Each of these sign variants might be traced back to different PRIs by the described procedure.
Another idea is to take the results of this procedure (b-values) and automatically fill gaps between attested areas so that the result is one large area of use on the map rather than several isolated colored counties. This could be done on the basis of nearness of neighboring areas surrounded by attested areas. For this completion procedure competing forms (different regional variants used for the same concept) should be taken into account: When a presumed area of use is to be extended to a non-attested area on the basis of geographical nearness this should only be done when this area is not attested for another competing sign.

6.4 Lexicographical Perspective

In sign language variation studies regional distribution of lexical variants usually has been dealt with by taking sites or predefined regions as a starting point and collecting data to determine which signs are used for certain concepts there. Then results can be compared with regard to number of variants and subvariants and the overlap of use in the different regions can be investigated. Regions have been defined on grounds of presumed or known differences within the language communities, small pilot studies or presumed or known influences of different locations of residential schools. The point here is, that usually the analysis looks at predefined regions and the use of signs therein.
In this study, the direction of focus has been turned around to facilitate a lexicographical perspective on regional distribution. The individual sign is the starting point of the analysis and the target of investigation is where exactly this particular sign is being used. This can be done without relying on predefined larger dialectal areas. The corpus data can speak for itself. It reveals the relevant areas of use for each sign through distributional maps produced directly from the corpus. This type of information is useful when writing a lexicographical description of signs in dictionary entries.

6.5 Dialectal Regions

The geographical boundaries between areas of use of different regional lexical variants for the same concept are called isoglosses. Corresponding isoglosses of several sets of signs with similar distributional patterns can be taken as indications of boundaries of dialectal regions. This is not only the case for lexical variants but also for all kinds of linguistic variables that display comparable patterns of regional distribution. Distributional maps cannot only be produced for the distribution of lexical variants but also for the distribution of other kinds of variation. The same procedure used here for the analysis of occurrences of signs can be adapted to occurrences of other phenomena coded and annotated in the corpus data.

6.6 Implications for Research on Color Signs

The elicitation of colors in the task *elicitation of isolated signs* was designed to gain data on lexical variation across regions, it was not intended to for a study on basic color terms in DGS. With the exception of one color-blind informant all informants were able to spontaneously give their color signs, some of them showed more than one variant (which were all included in the study). In few cases informants were unsure about the color presented, in three cases informants misinterpreted orange for beige. This might be due to the selection of the particular color as stimulus, lightning conditions at the site or the vision of the informants. The very high number of tokens for black (cf. table 1) can be explained by the elicitation setting. A black screen was used to elicit the color black and at the end of the task a black screen appeared to signal the end of a task in the same way as in other tasks. Most informants reacted to this black screen showing their sign for black again. Only in few cases an informant used the same manual sign form with different mouthings to name different colors. The most commonly used sign was RED1, which was used by almost all signers across the country with very few exceptions. For black (2 main variants), purple (3 main variants) and white (4 main variants) only few stable variants were found while a high number of variants were found for grey, green, brown and yellow. Some signs were used for more than one color.
There does not exist one single set of color signs for DGS as a whole. The observed high variation and complex distributional patterns of signs for colors in DGS might present a challenge for the research on basic color terms at the present state of research. Several combinations of regional variants that overlap to various degrees have to be taken into account for future studies on color signs.

7. Conclusion

The preliminary analysis of regional distribution of color signs from the DGS Corpus is one example of the many ways an annotated corpus can be utilized. Maps showing the regional distributions of tokens of sign variants can be generated directly from the annotations stored in a database together with lexical entries and relevant geo-

graphical data (metadata) on informants, as it is done in the iLex database and working environment. The visualization of the data on a geographical map provides a quick overview on regional distribution and can thus support the annotation and lemma revision processes as well as be a valuable tool for describing signs and their use in dictionary entries. Naturally, the results of such visualizations depend on the quality and consistency of the annotations and the existence of relevant geographical metadata on informants. First analyses of the signs for colors confirms the expectation that in DGS there is a high degree of variation in color signs and that a certain extent of these variants can be shown to be regional variants.

8. Acknowledgements

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.

9. References

Hanke, T. (2002). iLex. A tool for Sign Language Lexicography and Corpus Analysis. In González Rodriguez, M. & Paz Suarez Araujo, C. (Eds.), *Proceedings of the Third International Conference on Language Resources and Evaluation. Las Palmas de Gran Canaria, Spain*; Vol. III. Paris: ELRA, pp. 923-926.

Hanke, T., Storz, J. (2008). iLex - A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In Crasborn, O., Efthimiou, E., Hanke, T., Thoutenhoofd, E.D., Zwitserlood, I. (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, pp. 64-67. [Online available: http://www.lrec-conf.org/proceedings/lrec2008/workshops/W25_Proceedings.pdf].

Hollman, L., Sutrop, U. (2010). Basic Color Terms in Estonian Sign Language. In *Sign Language Studies*, 11(2), pp. 130-157.

Johnston, T.A. (2003). Language Standardization and Signed Language Dictionaries. In *Sign Language Studies*, 3(4), pp. 431-468.

König, S., Konrad, R., Langer, G. (2008). What's in a sign? Theoretical Lessions from Practical Sign Language Lexicography. In J. Quer, (Ed.), *Signs of the Time. Selected Papers from TISLR 2004*. Hamburg: Signum, pp. 379-404.

König, S., Konrad, R., Langer, G., Nishio, R. (2010). How Much Top-Down and Bottom-Up do We Need to Build a Lemmatized Corpus? Poster presented at the *Theoretical Issues in Sign Language Research Conference (TISLR 10), Sept 30 - Oct 2, 2010 at Purdue University, Indiana, USA*. [Online available: http://www.purdue.edu/tislr10/pdfs/Konig Konrad Langer Nishio.pdf].

Konrad, R. (2011). *Die Erstellung von Fachgebärdenlexika am Institut für Deutsche Gebärdensprache (IDGS) der Universität Hamburg (1993-2010).* Corrected Version. [Online available: http://www.sign-lang.uni-hamburg.de/projekte/mfl/konrad_2011_fachgeblexika.pdf].

Konrad, R., Hanke, T., König, S., Langer, G., Matthes, S., Nishio, R., Regen, A. (2012). From form to function. A database approach to handle lexicon building and spotting token forms in sign languages; this issue.

Lucas, C., Bayley, R., Valli, C. (2001). *Sociolinguistic Variation in American Sign Language*. Washington, D.C.: Gallaudet Univ. Press.

McKee, R., McKee, D. (2011). Old Signs, New Signs, Whose Signs? Sociolinguistic Variation in the NZSL Lexicon. In *Sign Language Studies*, 11(4), pp. 485-527.

McKee, D., McKee, R., Major, G. (2008). Sociolinguistic Variation in NZSL Numerals. In de Quadros, R.M. (Ed.), *Theoretical Issues in Sign Language Research 9. Sign Languages: spinning and unraveling the past, present and future. TISLR9, forty five papers and three posters from the 9th Theoretical Issues in Sign Language Research Conference, Florianopolis, Brazil, December 2006*, pp. 296-313.

Nishio, R., Sung-Eun Hong, S.-E., König, S., Konrad, R., Langer, G., Hanke, T., Rathmann, C. (2010). Elicitation methods in the DGS (German Sign Language) Corpus Project. In *Workshop Proceedings: 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies. Language Resources and Evaluation Conference (LREC), Valletta, Malta, May 22.-23.05.2010*, pp. 178-185. [Online available: http://www.sign-lang.uni-hamburg.de/lrec2010/lrec_cslt_01.pdf].

Schembri, A., McKee, D., McKee, R., Pivac, S., Johnston, T., Goswell, D. (2009). Phonological variation and change in Australian and New Zealand Sign Languages: The location variable. In *Language Variation and Change*, 21, pp. 193-231.

Schermer, T. (2003). From Variant to Standard: An Overview of the Standardization Process of the Lexicon of the Netherlands over Two Decades. In *Sign Language Studies*, 3(4), pp. 469-485.