Towards a Visual Sign Language Corpus Linguistics

Thomas Hanke
Institute of German Sign Language and Communication of the Deaf, University of Hamburg, Germany
E-mail: thomas.hanke@sign-lang.uni-hamburg.de

Abstract

Visualisations have a long tradition in linguistics, as in many fields dealing with complex structure. New forms of representations have been introduced to Visual Linguistics in the recent past, e.g. to help the researcher find the needle in a haystack, i.e. corpus. Here we present visualisation services available in iLex making a combined corpus and lexical database visually accessible. While many approaches suggested for textual languages transfer to sign language data as well, others explore sign-specific structure, such as multi-dimensional concordances not being restricted to sequentiality. Experimental combinations of animated visualisation and image processing might support the researcher to compensate for incomplete high-quality (=manual) annotation. In the long run, we see the potential that visualisation and data manipulation go hand in hand, allowing future user interfaces that are less text-heavy than today's sign language annotation environments.

Keywords: annotation, lexical database, iLex, visualisation, SQL, charting, mapping, geospatial data, R, graphviz, D3js

1. Introduction

Even though sign linguistics works on a visual language, it is not visual itself, or not more than linguistics on any other language. As Visual Linguistics, by no means being a new field, but having received increased attention over the last years, often operates on the levels of types and their relations to token, informants, or other types, many of the visualisation ideas transfer one-to-one to sign language corpora. With many interesting approaches having appeared in the last years, there is a lot to gain from.

Whether the researcher tries to find the needle in a haystack (like interesting structure worth a closer look or potential encoding errors maybe showing as either clusters or outliers in a visual representation) or to get an intuition what hypotheses to formulate and test, visualisation techniques should be readily available and well integrated into the sign language corpus linguist's workflow.

Visual representations are also most useful when illustrating complex relations to others, be it colleagues or students. Depending on the audience, the same representations that are used by the researcher to get an overview or detect new facts may be used, or more sophisticated graphics need to be produced, often abstracting further away from the actual data.

In addition to the degree of sophistication (from quick & dirty to fine-tuned graphics for lectures, for example), another independent dimension has come up in the past years: The degree of interactiveness. For visual representations that go into traditional print publications, non-interactive graphics are enough. For slides, more and more researchers make use of interactive visual representations. Animation not only makes visuals more attractive, but also allows the presenter to direct the viewer's (or even user's) attention to specific aspects. But interactive graphics also make sense for the researcher him-/herself when it opens the possibility to sort, zoom, or focus the attention or move back and forth on a timeline. With modern libraries such as D3js making this kind of display easier to implement, more and more researchers want to explore the potential of such displays.

2. Data Visualisation in iLex

As iLex is a corpus and lexical database (cf. Hanke, 2002, and Hanke & Storz, 2008), providing the data is "simply" a question of selection. We use SQL queries to provide the data to be rendered since SQL is a very powerful way of searching, selecting, grouping and ordering the data, spanning annotation and lexical database. The obvious disadvantage of this approach is that the user needs a good command of SQL to produce the tabular data s/he is interested in. To partially overcome this problem, iLex allows the user to store "chart" definitions, i.e. the underlying SQL query as well as the chart style. That way, the user him-/herself or any other user can execute the same chart at a later point of time, either on the same data or on other data points of the same category. iLex charts are either global or take data points of a certain category as input, like the types selected in a list of types. Thereby, it is easy to produce graphics specific for a set of types (or concepts etc.) the user is interested in – without having to read or even understand the SQL. This holds true for all kinds of charts implemented in iLex:

2.1. Business Charts in iLex (figs. 1-4)

iLex can convert tabular data into pie charts, bar charts or scatter plots most users are familiar with from popular spreadsheet applications. While the customisation options fall short compared to specialised application, the user can select in iLex which data points to create the graphics for, there is no need to copy the data elsewhere, and most importantly the user can double-click on a bar or pie segment or scatter point to open the related data point, or, in case of aggregation, a list showing all data points belonging to the selected aggregate.

Typical uses include token counts for selected types grouped by informant metadata such as sex and age group but also statistical data on annotation progress.

2.2. Graphs, Nets and Lattices in iLex (fig. 5)

Types and concepts quite naturally form complex nets that can be visualised inside iLex by virtue of the Graphviz library[1] integrated (cf. Gansner & North, 2000). Graphviz implements several algorithms to layout complex graphs with minimal edge overlaps. Double-clicking nodes or edges may open relevant detail.

2.3. Maps in iLex (figs. 6-7)

The combination of corpus data and related informant metadata allows for most interesting geolinguistic queries, such as the regions where users of a particular sign are from when trying to make up one's mind about the regional distribution of signs. Such data obviously is best displayed in maps.

iLex makes use of sophisticated geospatial R scripts [2] to plot the maps (Perpiñán Lamigueiro, 2014). For this to work, some data is needed in the background that relates geographical regions of interest to regions on a map. This data needs to be preloaded into iLex to match the regional distribution of target countries the database contents is related to. Again, the chart definition determines what happens when the user double-clicks on a map tile.

In our database, we offer geolinguistic queries on different levels of granularity (states, counties and data collection regions of the DGS Corpus[3] project[4]). While the infrastructure would also allow maps showing the exact living places of informants using a specific sign, such queries are generally blocked in this database for data privacy reasons as with a rather small set of informants from a regionally distributed minority re-identification often is possible from the living place alone.

2.4. Interactive Graphs in iLex (figs. 8-10)

The most recent addition to iLex's charting capabilities is the integration of D3js[5], a JavaScript library to design graphics that have more interactive functions than the aforementioned chart types (cf. Murray, 2013). There is a plethora of business chart and graph styles available building on D3js[6], but any real application requires tweeking the JavaScript code so that some JavaScript programming skills are needed to integrate new styles into iLex.

The advantage of these interactive graphs is that one can program them in a way to display a node's children when double-clicked or just grab a node and move it to another part of the window e.g. to sort by individual criteria. The logical next step would be to use these graphs not only for visualising data, but also for manipulating them. D3js has all the needed capabilities and easily connects with the iLex database. So in the long run, we expect such graphs to replace the text-heavy tabular data displays used all over the place. We hope that over time D3js will develop in a way to allow a clear separation of display and manipulation code so that security measures can apply. For the time being, we do not allow data manipulation SQL code inside D3js code, but only queries.

2.5. Exporting Charts from iLex

All charts created in iLex are in Scalable Vector Graphics format (svg) and thus can directly be integrated into web pages. For other programs not capable of importing svg, iLex allows printing the chart to PDF. Unfortunately, only the final view of animated visualisations shows in the PDF. So for exporting into slides etc., there also exists the option to export to a movie file.

3. Relations Explorable via Charts

A lexical database has a rich inventory of explicit relations between entities that can be visualised in a task-specific way. In addition, any distance measure defined between types implicitly establishes additional relations between them. We here explore similarity of HamNoSys descriptions; ASL-LEX (Sevcikova Sehyr et al., 2016) demonstrates that the same is possible and extremely insightful, based on phonological properties.

Combing corpus and lexical data in the database, there are both traditional and sign-specific approaches: With sign languages being able to articulate two (one-handed) signs at the same time, concordances become multi-dimensional. For our purposes, a concordance graph that color-codes the dimensions seems to be a good solution.

With the corpus data providing durations of tokens in a reliable way, it becomes possible to observe a signer's signing speed over the course of conversations in different elicitation settings.

While it is possible to combine various chart types into one window using the R and D3js renderers, e.g. to have pie charts for each region shown in a map, iLex offers another easy and flexible way of combining charts: The hyperlink determining what happens when the user double-clicks a chart segment, a graph node, or a map tile can also refer to another chart. That way, cascades of charts can be built with minimal effort. For example, the user can select from types visually grouped by phonetic features in order to see their regional distribution on a map.

[1] http://www.graphviz.org ; last access: March 26, 2016

[2] http://www.r-project.org ; last access: March 26, 2016

[3] http://dgs-korpus.de ; last access: March 26, 2016

[4] The maps have been produced from data provided by a German public body responsible for geodata, cf. http://www.geodatenzentrum.de/docpdf/vg1000.pdf ; last access: March 26, 2016

[5] http://www.d3js.org ; last access: March 26, 2016

[6] http://bl.ocks.org ; last access: March 26, 2016

4. Animation Overlays to Videos (fig. 11)

While not technically being a "chart" type in iLex, graphics video overlays look like an animation when the video is played back, thereby establishing an interesting visualization per se: One type of annotation that iLex offers is coordinates (of a point or rectangle, measured in percentages of the video resolution width and height). This is typically used to make the results from automatic 2D face and hand tracking available to the annotator as these points and rectangles tags can be superimposed to the video. Transparency ramp functions are one possibility to achieve a ghosting effect to the animated rectangles and points. When combined with a grid of positions to memorize, this results in "temporal heat map" that at least visually comes close to Dalle's idea to model signing space (cf. Braffort & Dalle, 2007).

5. Future Developments

While we expect many more chart styles being used for sign language corpus work with the technology available, it remains a goal for us to make their definition easier, by providing a query language or a query builder tool that is closer to linguistics than SQL: For written languages, there are a number of impressive examples such as ANNIS (cf. Krause & Zeldes, 2014).

Video overlays and the underlying annotation are currently restricted to two-dimensional video coordinates. This means that annotations do not transfer from one camera perspective to another which is rather annoying for the annotator. Our plan is to make three-dimensional structure annotation available and feed 3D tracking data into it. Only then it will become to use this approach to verify manual annotation by also visualising the sign trace derived from the HamNoSys notation.

We are well aware that visualisations produced so far with the tools integrated into iLex are limited to linguistic categorisations of the signed texts that make up the content of sign language corpora. To explore the content itself in various humanities research dimensions, completely different approaches what to visualize may be needed (cf. Uboldi & Caviglia 2015).

6. 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.

The maps displayed in this report are based on data under German federal government copyright: © GeoBasis-DE / BKG 2013 (data modified).

7. References

Braffort, A., Dalle, P. (2007). Sign Language Applications: Preliminary Modeling. *Universal Access in the Information Society*, 6(4), pp. 393--404.

Gansner, E., North, S. (2000). An open graph visualization system and its applications to software engineering. In *Software – Practice and Experi*ence, (30)11: 1203—1233.

Hanke, T. (2002). iLex. A tool for Sign Language Lexicography and Corpus Analysis. In M. González Rodriguez, & C. Paz Suarez Araujo (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. [Online resource; URL: http://www.lrec-conf.org/proceedings/lrec2002/pdf/330.pdf ; last access: March 22, 2016]

Hanke, T., Storz, J. (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood, & E. Thoutenhoofd (Eds.), *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages.* Paris: ELRA, pp. 64--67.

Krause, T., Zeldes, A. (2014): ANNIS3: A new architecture for generic corpus query and visualization. In *Digital Scholarship in the Humanities 2014*.

Murray, S. (2013). *Interactive Data Visualization for the Web*. O'Reilly: Sebastopol CA.

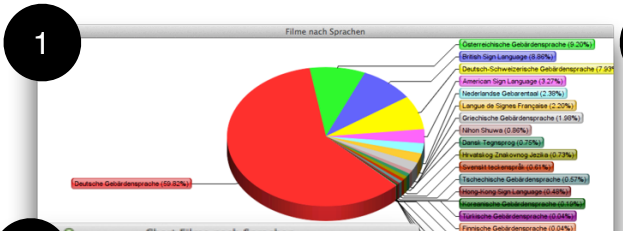
Perpiñán Lamigueiro, O. (2014). *Displaying Time Series, Spatial, and Space-Time Data with R*. Taylor & Francis. Hoboken.

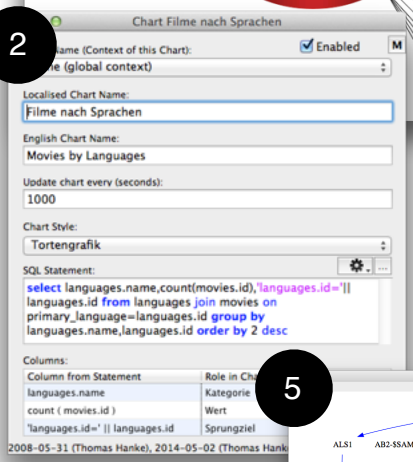
Sevcikova Sehyr, Z., Caselli, N., Cohen-Goldberg, A., Emmorey, K. (2016). ASL-Lex, a Lexical Database for American Sign Language. Poster presented at the Theoretical Issues on Sign Language Research (TISLR12) Conference in Melbourne, Australia, 4-7 Jan 2016. [Online resource; URL: http://slhs.sdsu.edu/llcn/files/2016/01/TISLR2016_ASLLEX_FINAL.pdf ; last access: March 26, 2016]

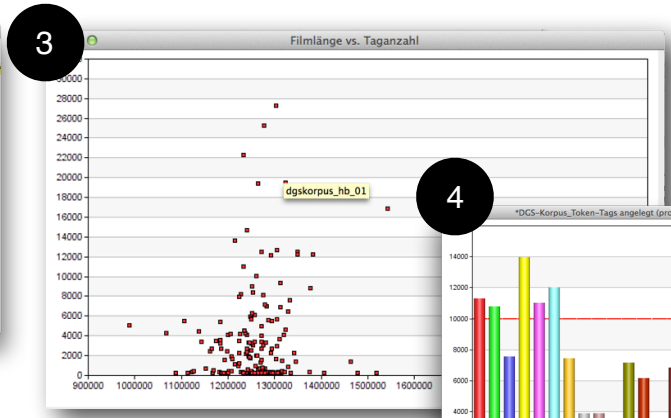
Uboldi, G., Caviglia, G. (2015). Information Visualizations and Interfaces in the Humanities. In Bihanic, D. (Ed.), *New Challenges for Data Design*. Springer: London, pp. 207—218.

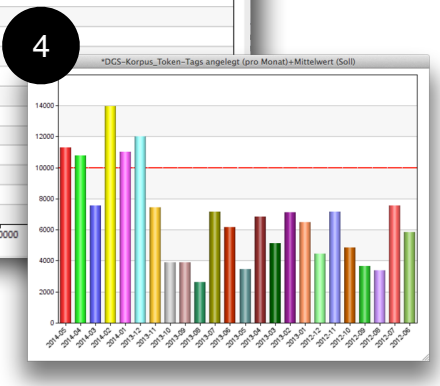
8. Figures Legend

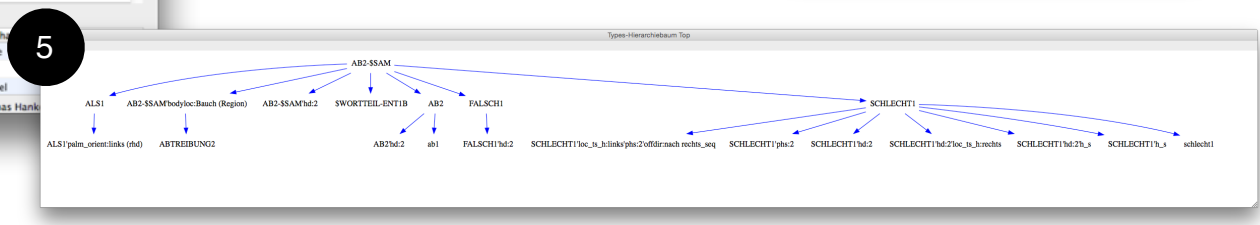
Fig. 1: Pie chart, fig. 2 is the corresponding chart definition: Distribution of movies by language. Figs. 3 and 4: Scatter chart and bar chart on progress monitoring. Fig. 5: Type hierarchy. Fig. 6: Data collection (sub-) regions with informants using FRAU2, one of several signs meaning woman. Fig. 7: Regional distribution of informants in the DGS corpus project. Fig. 8 Force-directed graph showing a segment of the type hierarchy around FRAU2. Fig. 9 Chord graph showing the distribution of source and goal in directed verb by token counts. Fig. 10: Excerpt of a syntax diagram for the HamNoSys notation for the sign AB1A (away). Fig. 11: Rectangle annotation overlayed to video.

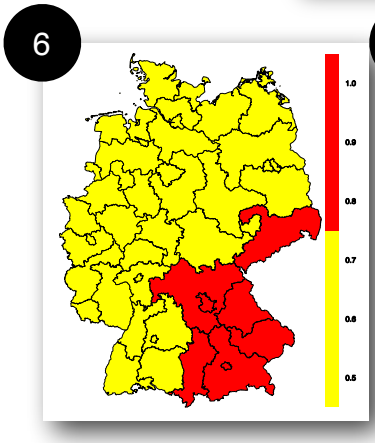

1
Filme nach Sprachen

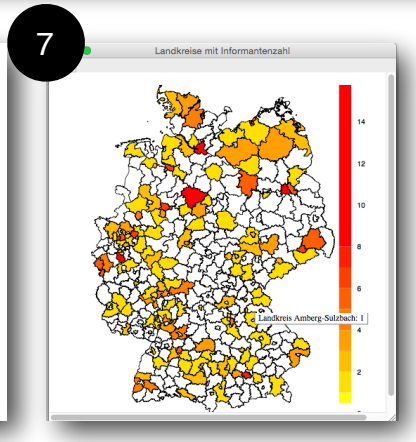

2
Chart Filme nach Sprachen
Enabled
Localised Chart Name:
Filme nach Sprachen
English Chart Name:
Movies by Languages
Update chart every (seconds):
1000
Chart Style:
Tortengrafik
SQL Statement:
select languages.name,count(movies.id),'languages.id='||
languages.id from languages join movies on
primary_language=languages.id group by
languages.name,languages.id order by 2 desc
Columns:
Column from Statement
languages.name
count (movies.id)
'languages.id=' || languages.id

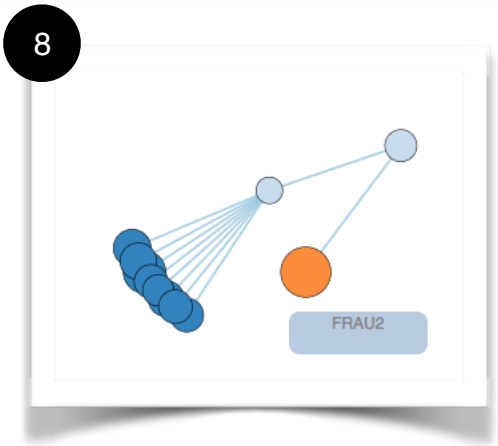

3
Filmlänge vs. Taganzahl
dgskorpus_hb_01

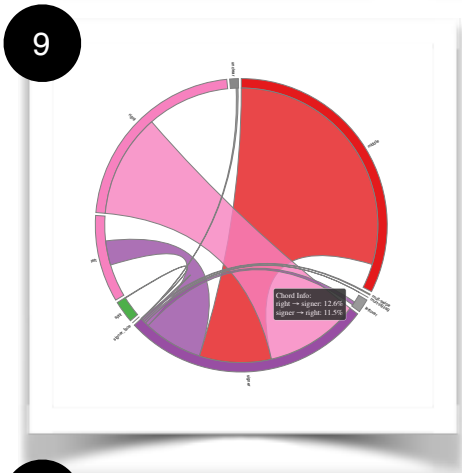

4
*DGS-Korpus_Token-Tags angelegt (pro Monat)+Mittelwert (Soll)

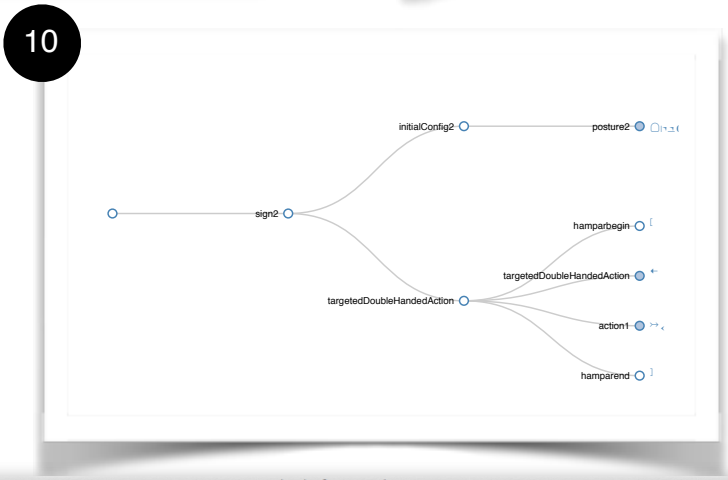

5
Types-Hierarchiebaum Top
AB2-$SAM
ALS1
AB2-$SAM'bodyloc:Bauch (Region)
AB2-$SAM'hd:2
SWORTTEIL-ENT1B
AB2
FALSCH1
SCHLECHT1
ALS1'palm_orient:links (rhd)
ABTREIBUNG2
AB2'hd:2
ab1
FALSCH1'hd:2
SCHLECHT1'loc_ts_h:links'phs:2'offdir:nach rechts_seq
SCHLECHT1'phs:2
SCHLECHT1'hd:2
SCHLECHT1'hd:2'loc_ts_h:rechts
SCHLECHT1'hd:2h_s
SCHLECHT1'h_s
schlecht1

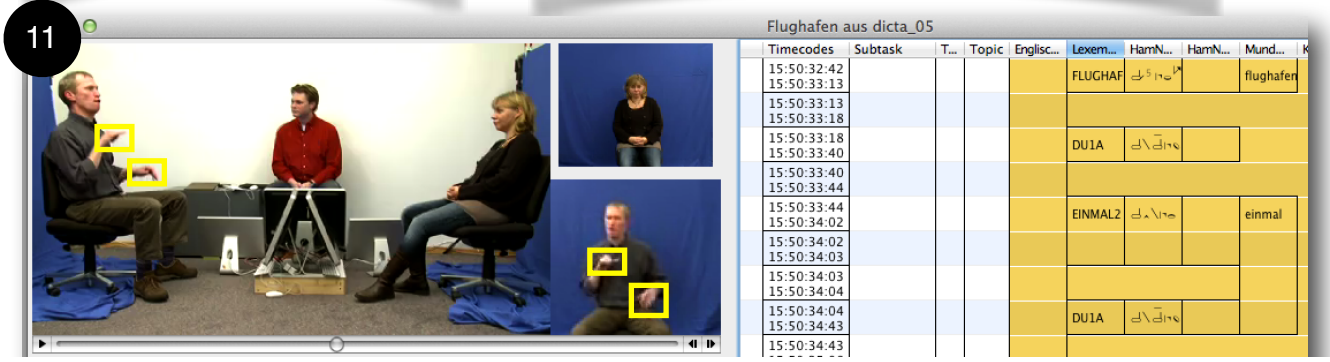

6


7
Landkreise mit Informantenzahl
Landkreis Amberg-Sulzbach: 1


8
FRAU2


9


10
initialConfig2
posture2
sign2
hamparbegin
targetedDoubleHandedAction
targetedDoubleHandedAction
action1
hamparend


11
Flughafen aus dicta_05
Timecodes
Subtask
Topic
Lexem...
HamN...
Mund...
FLUGHAF
flughafen
DU1A
EINMAL2
einmal
DU1A