Visualising Language in Space - New Approaches in Linguistic Cartography

Christina Mutter and Florian Zacherl

Abstract

This short paper describes the current approaches the Digital Humanities project VerbaAlpina of Munich University takes to visualise linguistic data in the Alpine region. Whereas traditional linguistic cartography has several disadvantages, VerbaAlpina manages to address these by the development of an interactive map which is at the heart of an innovative web-based research environment. Among other things, one of the main achievements of this interactive map is the integration of so far coexisting cartographic traditions. To create a trustworthy visualisation the user is presented with multiple filters on the full data set, different displaying methods and an in-depth view into the underlying utterances given by the informants.

Index Terms: Human-centered computing—Visualization—Visualization techniques—Geographic visualization; Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Web-based interaction

1 Introduction

In dialectology space is one of the most important dimensions of linguistic variation, from national territories down to regional or even local dialects, as those of particular city quarters. Visualisation is therefore an important part of the process of analysing linguistic variation, since the visual representation of spatial relationships often provides information about other types of relationships. For example, a spatial relationship can be interpreted as a temporal relationship, i.e. as a diachronic interpretation of a diatopic representation. It can also be interpreted as a similarity relationship, e.g. in etymological or semantic terms. In this way, spatial proximity can draw attention to similarities that might otherwise not be recognizable. Hence, visualisation can make similarities or differences and, as a result, cultural and linguistic-historical connections between linguistic data more evident [6, 486/7].

2 Pros and Cons of traditional linguistic cartography from the perspective of a linguist

The common visualisation of geolinguistic variation is mapping. Nevertheless, geolinguistics is far from using standardized forms of cartography. In fact, two main traditions are coexisting since the beginning of dialectology in the 19th century, with pros and cons each of them: the so-called analytic maps (I) and the synthetic ones (II).

(I) Analytic maps which are typical for the Romance tradition give a full representation of every utterance given by the informants, like for example in the atlas “Sprach- und Sachatlas Italiens und der Südschweiz” (AIS) [7]. This form of visualisation has the advantage that it offers very detailed information which is reliable to the source, traceable and verifiable. On the other hand, analytic maps are, however, comparatively confusing and unclear. The focus is here on the documentation of the single utterances and the readers need to understand the spatial relationships between the documented language forms by themselves.

(II) Synthetic maps which represent the Germanic tradition, like the maps of the atlas “Vorarlberger Sprachatlas” (VALTS) [5], do not provide entire linguistic forms but abstract graphic representations of certain linguistic features in a qualitative, quantitative or probabilistic way. This has the advantage that these maps are clearly arranged so that the spatial relations between single utterances become directly apparent. However, synthetic maps are difficult to verify and lack transparency [2], [8], [11].

Furthermore, in traditional linguistic cartography vocabulary is viewed exclusively from an onomasiological perspective[1] and traditional linguistic atlases usually provide only a monolingual view on certain dialect regions. However, this does not allow to understand whether a certain linguistic characteristic is only used in one dialect or one language or whether linguistic phenomena cross dialectal, linguistic or even political borders. Another disadvantage of traditional linguistic cartography consists in the fact that it depends on time and place since traditional linguistic atlases are not available everywhere and at all times. Despite the numerous disadvantages, however, it cannot be denied that traditional linguistic cartography has the clear advantage that these works are permanent and can be unambiguously referenced.

3 The added value of the interactive map of the research project VerbaAlpina

In order to address the disadvantages of traditional cartography described in chapter 2, first of all mapping tools are required that make it possible to integrate the two before-mentioned visualisation traditions. Such a mapping tool, for example, is the interactive map of the research project VerbaAlpina[2], which is at the heart of the innovative lexicographic online platform developed by the

[1] Onomasiology is the study of designations and aims at finding words that describe a given concept. The onomasiological approach answers the question “How do you express X?”. The opposite approach is the semasiological one which asks what concepts a word refers to. [16]

[2] VerbaAlpina (`https://www.verba-alpina.gwi.uni-muenchen.de`) is a long-term research project of Munich University, which has been funded by the German Research Foundation (DFG) since October 2014. The project investigates the Alpine lexis of three conceptual domains in the Alpine region where dialects and languages belonging to three large language families (Germanic, Romance and Slavonic) are spoken. VerbaAlpina is a cooperation of the Institute of Romance Philology and the LMU Center of Digital Humanities of Munich University (`http://www.itg.lmu.de`) and offers a cross-disciplinary approach by combining linguistics, ethnology and information technology. The data collection of the three conceptual domains is broken down in three stages. Stage one (from October 2014 to October 2017) focused on vocabulary related to alpine pasture farming, in particular, milk processing. In the current phase (from November 2017 to November 2020) the project is concerned with the lexis of the domains fauna, flora, landscape formation and weather. The last stage (from December 2020 to December 2023) will focus on the vocabulary of modern alpine life (ecology, tourism). The majority of the linguistic data VerbaAlpina gathers and analyses derived from linguistic atlases and from geo-referenced dictionaries from the past one hundred years. This data is supplemented by current dialect words via crowdsourcing.

project. The area under investigation of VerbaAlpina is the Alpine region[3], which is highly fragmented with regard to languages and dialects. The project investigates selectively, i.e. limited to three conceptual domains, and analytically the Alpine region in its cultural and historical linguistic unity. The web-based research environment of VerbaAlpina makes it possible to integrate the two visualisation traditions. Thus, on the interactive map of VerbaAlpina at first glance ”synthetic” maps are displayed but at second glance, after clicking on the individual symbols on the map, one also gets access to the single utterances. In this way also empirical transparency is guaranteed [8].

In addition, the web-based research environment of VerbaAlpina simplifies a cross-national and cross-linguistic investigation and at the same time it overcomes the restriction of traditional geolinguistics to political units (nation states). The visualisation of linguistic data via the interactive map also makes it possible to integrate different data sources. Thus, not only data from linguistic atlases is displayed but also data from dictionaries in the Alpine region as well as data collected via crowdsourcing. Hence, vocabulary can not only be viewed from an onomasiological perspective (as in traditional geolinguistics) but also from a semasiological one using appropriate filters on the interactive map. The visualisation of linguistic data can always be carried out qualitatively or quantitatively and based on geographic polygons or abstract hexagons (cf. chapter 4). Furthermore, the interactive map offers the function to create so-called synoptic (i.e. collective) maps. With this function the user has the possibility to capture an individual selection of data on a synoptic combination map. In this way, the range and distribution of any linguistic and non-linguistic feature can be visualised in context. The respective selection can be saved and recalled at a later point in time [10]. Compared to traditional linguistic atlases, a decisive asset of the interactive map is that it is independent of time and place and can be accessed anytime and anywhere thanks to its online format. However, one of the biggest challenges in dealing with web-based research environments consists currently in the durability of the data. This is challenging since online data should meet certain criteria, defined by the so-called FAIR principles, and thus meet the postulates ”findable”, ”accessible”, “interoperable” and “reuseable” [17], [13]. Currently, a relatively large effort is required to achieve this, as appropriate approaches are still being developed.

4 Finding appropriate visualisations for linguistic data

Although all VerbaAlpina modules principally can be used by a broad public the main focus lies on a scientific audience. Therefore an important consideration regarding the visualisation is to make it appealing to traditional linguists.

One aspect is to create a certain familiarity by following well-known representations. If the user first selects a concept or lexeme from the drop-down menus or via the search function, a traditional synthetic map with point symbols and an appropriate legend is shown. This starting point enables an easy habituation for scientists that are potentially unaccustomed with this type of online resource. A natural next step can be to combine the currently chosen data set with other linguistic or exta-linguistic data.

The type of visualisation depends heavily on context, also it might not be obvious which one works well for a given data set. Another factor may be certain user preferences. VerbaAlpina addresses this problem by offering different methods of visualising the same data. As mentioned before, the default view is a classical point symbol map, but it is easily possible to switch to a quantitative view to help find patterns and maximums, especially in confusing situations with many symbols. For that, different polygon layers, mostly administrative borders (communities, regions, nations etc.), can be coloured in the fashion of a heatmap. Another option is to switch to the hexagon mode that replaces the geographic polygon layers by equally sized hexagons that are roughly arranged like their real-word equivalents. This is particularly useful in the alpine areas where large communities are oftentimes very sparsely populated while larger settlements with many inhabitants appear relatively small. Figure 1 shows different views for the same data points.

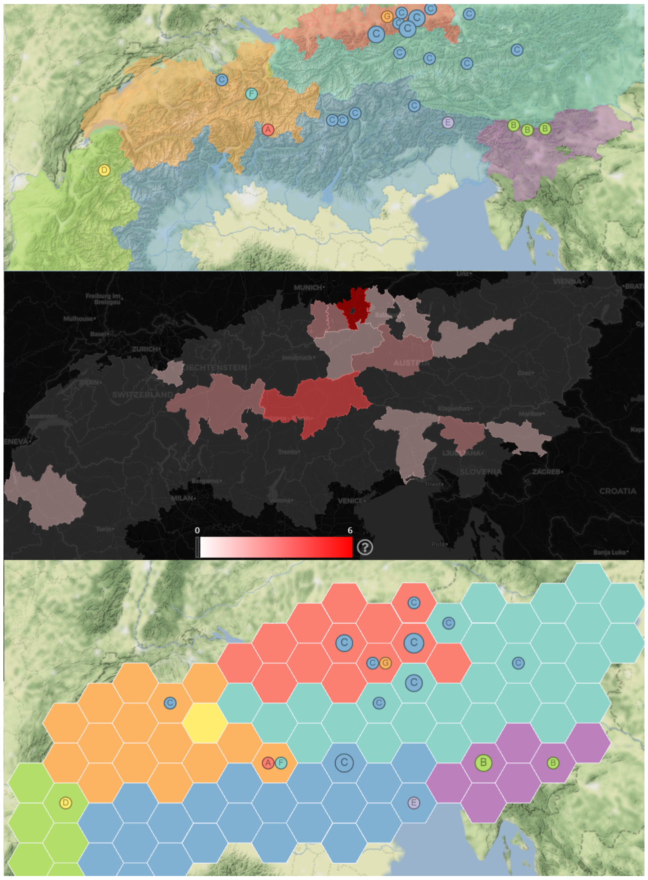


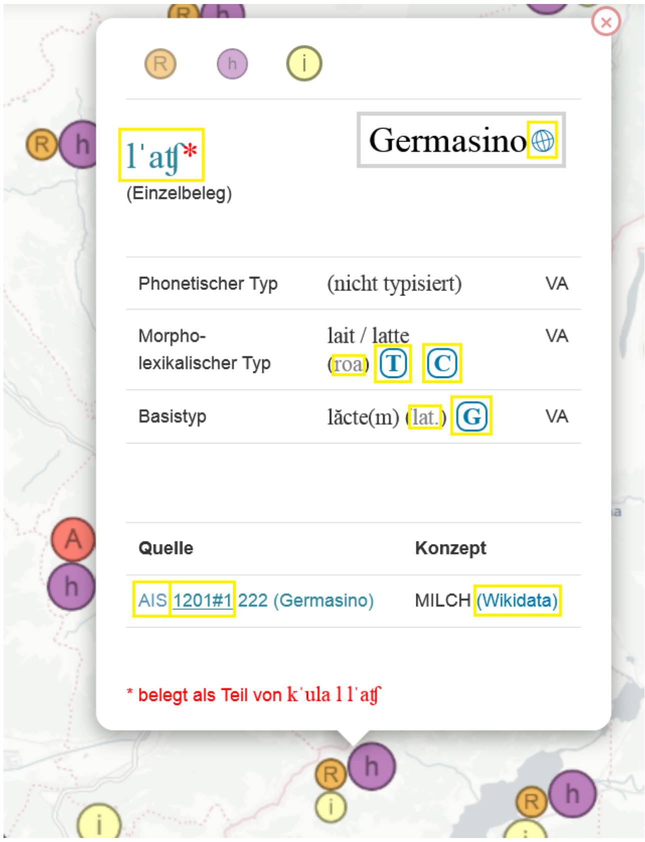
Figure 1: Different forms of visualisation: Geographic map with point symbols (top), heat map (middle) and hexagon map (bottom)

As mentioned in chapter 2, one main disadvantage of classical synthetic maps is the lack of transparency since the underlying raw data can not be accessed. Modern web technology makes it considerably easier to solve this problem by interactivity. In principle, this corresponds to the so-called visual information seeking Mantra: “overview first, zoom and filter, then details on demand” [15]. Although at first an aggregated overview is shown, it is possible for each data point to access the full linguistic and meta-information including the source from which the specific utterance was taken. Figure 2 shows the detail screen for one specific utterance taken from the linguistic atlas AIS [7].

Many elements on this screen give even more detailed information on user mouse interaction or are interlinked with other online resources, especially dictionaries and norm databases.

Naturally a project like VerbaAlpina, that has to merge data from multiple very different sources, needs some unification steps. One example is the conversion of all phonetic records given in vary-

[3] The area under investigation is limited to the territorial borders defined by the Alpine Convention, an international treaty between the Alpine Countries and the EU, for the sustainable development and protection of the Alps. This area covers a surface area of 190,700 km^2 and encompasses parts of six different countries (Austria, Italy, France, Switzerland, Germany and Slovenia) as well as two entire countries (Liechtenstein and Monaco). [1]



lˈaʧ*
(Einzelbeleg)

Germasino

Phonetischer Typ	(nicht typisiert)	VA
Morpho-lexikalischer Typ	lait / latte (roa) T C	VA
Basistyp	lăcte(m) (lat.) G	VA

Quelle	Konzept
AIS 1201#1 222 (Germasino)	MILCH (Wikidata)

* belegt als Teil von kˈula l lˈaʧ

Figure 2: The VerbaAlpina detail view for one specific utterance. All parts marked in yellow are interactive and/or linked to other online resources.

ing transcription systems into the International Phonetic Alphabet[4] (IPA). To keep this process as transparent as possible the user can get further information about the record encoding by hovering the transcription. In this example the record “l'aʧ” is labeled as encoded in IPA and the original transcription, in this case “láć” in the Böhmer-Ascoli convention, is given. If possible an online version of the source material is linked, so that the user can easily verify the given information.

All these aspects help to get an easy access to the interactive map, while opening up multiple different possibilities to work with the chosen subset of the VerbaAlpina data.

5 Technical background

From the very beginning the project VerbaAlpina was designed as an exclusively web-based project using open-source or free software as much as possible. An early design decision was to base its website on the content management system Wordpress[5]. With these premises the interactive map module is solely based on JavaScript and PHP with a MySQL [6] backend.

The basis for each visualisation (especially for different views on the same data set) is a sound data model. Many Digital Humanities projects utilise the xml-based TEI-format (**T**ext **E**ncoding **I**nitiative)[7] for data representation. As the words “text encoding” already hint, this is well-suited for representing text, but less ideal for encoding linguistic data from a variety of different sources. Therefore VerbaAlpina stores its data in a normalized (cf. e.g. [3]) relational database.

The principle idea of this model is to maintain independent lists of morphological types (lexemes) and concepts (meanings) which are connected by a “many-to-many” relation. Each of these relations is bound to a specific location and time. This opens up the possibility to create views in semasiological and onomasiological perspective as mentioned in chapter 3. Although this approach might seem intuitive from a data theoretical point of view, it differs significantly from a classical bilingual dictionary model in which lexemes are connected directly by their meaning. Lexemes whose meanings differ regionally, which are particularly interesting in dialectology, can only be properly represented with a more complex model.

Details of the VerbaAlpina data model can be found in its methodology section[8], especially in [12], [9] and [14].

The frontend uses the libraries Leaflet [9] for the basic map functionality and PixiJS[10], that utilises WebGL, for the map overlays. So the tool can highly efficient visualise even larger amounts of data. Details can be found in [4].

6 Conclusion

Summing up, the interactive map of VerbaAlpina solves the shortcomings of traditional linguistic cartography by merging and extending existing visualisation methods. Users are provided with the possibility to create and store their own compilations of data sets from different sources with the help of various tools and filters.

In spite of the several new options provided by modern web technology one major issue of web-based research environments remains the long-term accessibility of tools involved, especially regarding the compatibility of future software.

References

[1] Alpine Convention. Contracting parties. `https://www.alpconv.org/en/home/organization/contracting-parties/`. [Online; accessed 16-August-2019].

[2] P. Auer, J. Schmidt, and A. Lameli. *Language and Space. Vol. 2: Language Mapping. An international Handbook of Linguistic Variation*. De Gruyter, Berlin/New York, 2010.

[3] E. F. Codd. *The Relational Model for Database Management: Version 2*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1990.

[4] D. Englmeier. Technische Konzeption und Umsetzung der geografischen Datendarstellung. In *Methodologie*. VerbaAlpina-de 19/1, 2019.

[5] E. Gabriel. *Vorarlberger Sprachatlas mit Einschluss des Fürstentums Liechtenstein, Westtirols und des Allgäus*, vol. 1–5. Vorarlberger Landesbibliothek, Bregenz, 1985-2004.

[6] K. Gluth, M. Lompa, and H.-H. Smolka. Verfahren dialektologischer Karteninterpretation und ihre Reichweite. In W. Besch, U. Knoop, W. Putschke, and H. E. Wiegand, eds., *Dialektologie: ein Handbuch zur deutschen und allgemeinen Dialektforschung*. De Gruyter, Berlin/New York, 1982.

[7] K. Jaberg and J. Jud. *Sprach- und Sachatlas Italiens und der Südschweiz*. Zofingen, 1928-1940.

[8] T. Krefeld. Kartographie. In *Methodologie*. VerbaAlpina-de 19/1, 2019.

[9] T. Krefeld. Semantik. In *Methodologie*. VerbaAlpina-de 19/1, 2019.

[10] T. Krefeld. Synoptische Karte. In *Methodologie*. VerbaAlpina-de 19/1, 2019.

[11] S. Lücke. Einführung in die Geolinguistik (ITG/slu). In *Lehre in den Digital Humanities*. 1 ed., 2019.

[4] `https://www.internationalphoneticassociation.org/`

[5] `https://wordpress.org/`

[6] `https://www.mysql.com`

[7] `https://tei-c.org/`

[8] `https://doi.org/10.5282/verba-alpina?urlappend=%3Fpage_id%3D493`, currently available in French, German, Italian and Slovene

[9] `https://leafletjs.com/`

[10] `https://www.pixijs.com/`

[12] S. Lücke. Entity Relationship. In *Methodologie*. VerbaAlpina-de 19/1, 2019.

[13] S. Lücke. Fair-Prinzipien. In *Methodologie*. VerbaAlpina-de 19/1, 2019.

[14] S. Lücke and F. Zacherl. Mehrwortlexie. In *Methodologie*. VerbaAlpina-de 19/1, 2019.

[15] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pp. 336–343, Sep. 1996. doi: 10.1109/VL.1996.545307

[16] Wikipedia contributors. Onomasiology — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Onomasiology&oldid=807830470, 2017. [Online; accessed 16-August-2019].

[17] M. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. O. Bonino da Silva Santos, P. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. Evelo, R. Finkers, and B. Mons. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 03 2016. doi: 10.1038/sdata.2016.18