Eurographics Conference on Visualization (EuroVis) 2012
S. Bruckner, S. Miksch, and H. Pfister
(Guest Editors)

Volume 31 (2012), Number 3

The World's Languages Explorer: Visual Analysis of Language Features in Genealogical and Areal Contexts

Christian Rohrdantz[1], Michael Hund[1], Thomas Mayer[2,3], Bernhard Wälchli[4], Daniel A. Keim[1]

[1]Department of Computer Science, University of Konstanz, Germany
[2]Research Unit "Quantitative Language Comparison", LMU Munich, Germany
[3]Department of Linguistics, University of Konstanz, Germany
[4]Department of Linguistics, Stockholm University, Sweden

Abstract

This paper presents a novel Visual Analytics approach that helps linguistic researchers to explore the world's languages with respect to several important tasks: (1) The comparison of manually and automatically extracted language features across languages and within the context of language genealogy, (2) the exploration of interrelations among several of such features as well as their homogeneity and heterogeneity within subtrees of the language genealogy, and (3) the exploration of genealogical and areal influences on the features. We introduce the WORLD'S LANGUAGES EXPLORER*, which provides the required functionalities in one single Visual Analytics environment. Contributions are made for different parts of the system: We introduce an extended Sunburst visualization whose so-called feature-rings allow for a cross-comparison of a large number of features at once, within the hierarchical context of the language genealogy. We suggest a mapping of homogeneity measures to all levels of the hierarchy. In addition, we suggest an integration of information from the areal data space into the hierarchical data space. With our approach we bring Visual Analytics research to a new application field, namely Historical Comparative Linguistics, and Linguistic and Areal Typology. Finally, we provide evidence of the good performance of our system in this area through two application case studies conducted by domain experts.*

Categories and Subject Descriptors (according to ACM CCS):
Models and Principles [H.1.2]: User / Machine Systems—[Human information processing]
Information Interfaces and Presentation [H.5.2]: User Interfaces—[Graphical user interfaces]

1. Introduction

There are app. 6,900 modern natural human languages (see `http://www.ethnologue.com`), many of them endangered or moribund. The comparative analysis of the world's languages is a considerable challenge, which is traditionally addressed from three different sides. Historical-comparative linguistics deals with language families (genealogically related languages) which derive from largely homogeneous reconstructed proto-languages, such as Indo-European, through structural divergence in language change. Areal linguistics investigates how intensive language contacts seduce languages to converge structurally in linguistic areas such as South East Asia or Mesoamerica. Linguistic typology explores the full range of linguistic variability in terms of structural features, such as word order and number of grammatical cases. While typology traditionally tries to explain the distribution of structural features with other structural features, modern research has shown that linguistic diversity is not randomly distributed over the world, but that there are macro-areal patterns of continental or even hemispheric size [Dry92, Nic92, DH] which must be due to very old language contacts and/or genealogical relations that are not demonstrable with standard historical methods. This reunited the three disciplines in areal typology, which investigates typological, genealogical and areal properties in their interplay. [Dry92] divides the world into six regions (macro areas) where massive language contacts are most likely to have occurred. Wherever features in genera - genealogical units with a time depth of app. 3,500 years, such as Germanic or Romance - are not distributed the same way across all six regions this is taken as evidence for a non-random dis-

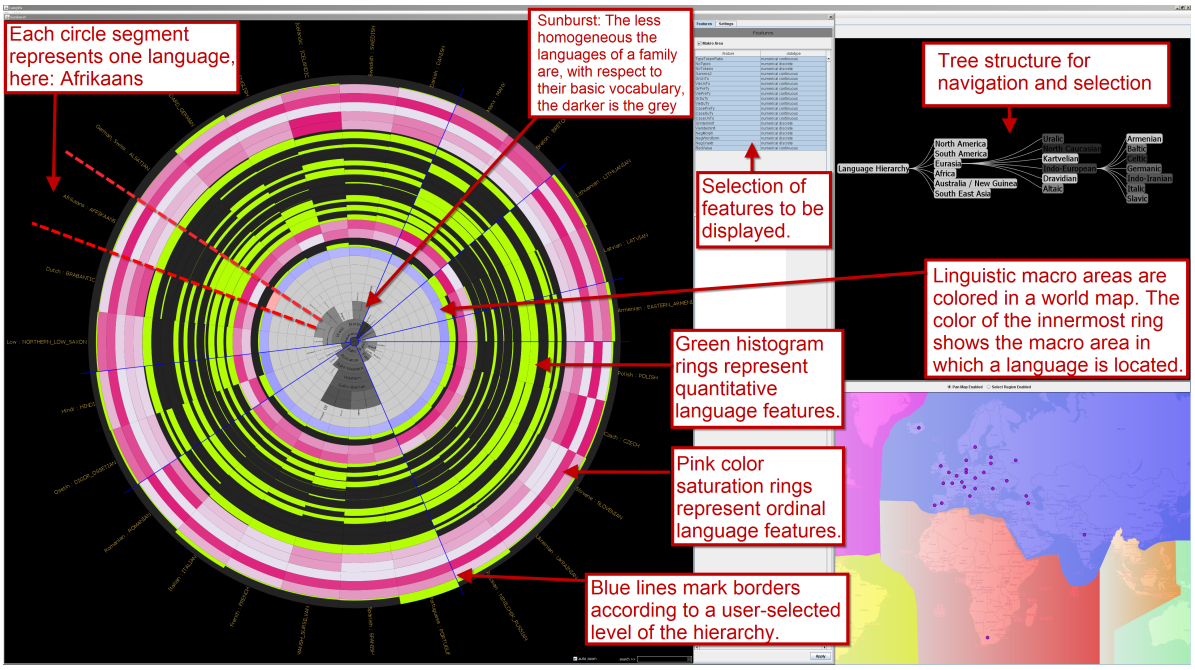


Figure 1: *High-resolution screenshot of the World's Languages Explorer showing the main components. The example shows for 27 Indo-European languages 19 language features that were automatically extracted from parallel Bible texts.*

tribution. Areal typology investigates among other things genealogical stability of features and their propensity to areal diffusion.

In recent years an increasing number of manually edited language data has been created, digitalized and made available to the public, some of which will be described in Section 2. The compilation of traditional typological databases, however, is expensive and time consuming, entailing many gaps and forcing typology to consider only a subset of the available data (typological sampling; [RB98]). An alternative is to extract typological features automatically from massively parallel texts (translations of the same text into many languages, such as the New Testament; [CW07]) with the advantage that languages can be compared with less data reduction directly on the level of language use rather than as idealized abstract systems.

Despite of the increasing availability of automatically and manually generated language features, until now, linguistic researchers have only marginally availed themselves of visualizations respectively advanced interactive visual interfaces for doing cross-linguistic comparisons and exploration. The very few examples include work on a phenomenon called vowel harmony, where vowel patterns in languages are represented through matrix visualizations [RMB*10]. The World Atlas of Language Structures (`http://wals.info`; see Section 2.2) offers a variety of language properties that are mapped to the geo-positions where the respective languages are spoken. Another approach combines a world map with other visual representations for the analysis of meaning evolution [TFES11]. Finally, the *Multitree* tool [Yps09] enables the user to visually access information about language relationships displayed as a node-link tree diagram. Yet, no work exists that combines both a geo-spatial and a hierarchical representation and would allow to visually compare multiple features at once.

In this paper, the goal is to provide a visual analytics system, the *World's Languages Explorer*, that enables the analysis of languages with respect to several research questions that domain experts have, such as: Are certain language features homogeneous within certain branches of the genealogy and diverse across different branches? This might be a trace of language change way back in history. Are there any outliers, that is, languages where a certain feature value surprisingly deviates from that of other closely related languages? If so, is this outlier value similar to that of other unrelated, but geographically close languages? This might point to a language change that was triggered by language contact, which is of special interest to linguists. More details are provided in Section 3.

This design study (see Figure 1) contains several contributions to the field of Visual Analytics: We propose to display the language genealogy as a Sunburst visualization and complement it with our *feature rings* which allow a cross-comparison of several features at once, within the hierarchi-

cal context of the language genealogy. Feature rings have different representations depending on whether they display quantitative, ordinal, or nominal features. Moreover, we suggest a mapping of homogeneity measures to all levels of the hierarchy. We also propose different means to integrate areal information into the hierarchical data space. A further contribution is that we bring Visual Analytics research to a new application field, namely Historical Comparative Linguistics, and Linguistic and Areal Typology.

The paper is structured as follows: In Section 2 we describe our automatic feature extraction and further data sources containing manually edited language features. Section 3 gives insight into the concrete tasks and requirements linguistic researcher have. Then, in Section 4, we give an overview on how this approach relates to other methods and techniques for visual data exploration. After that, in Section 5 we introduce our new system and give a detailed explanation of design decisions and our contributions. Section 6 next provides two application case studies showing real findings relevant to linguistic researchers. In Section 7 we discuss advantages and limitations of our approach and finally, we give a conclusion in Section 8.

2. Data

In this section we briefly describe our approach for automatic feature extraction (Section 2.1) as well as the external data sources used for our approach (Section 2.2).

2.1. Automated Linguistic Feature Extraction

Many simple linguistic features can be extracted from parallel texts, such as the New Testament, by indirect measurement [Juo08, Wäl12]. This holds especially for morphological typology. Morphological typology is a traditional field within linguistic typology concerned with assessing the degree of cross-linguistic variation in morphology, the internal structure of words [Gre60]. Five families of values — (i) degree of synthesis, (ii) amount of prefixing and suffixing, (iii) case, (iv) amount of internal inflection, and (v) synthetic vs. analytic negation marking — are extracted automatically from electronic parallel texts (here the Gospel according to Mark) in a diverse world-wide convenience sample of 161 languages and in 125 languages of Papua New Guinea (`http://www.pngscriptures.org`).

Languages differ in how much information is packed in a word (degree of synthesis). In parallel texts languages with more complex morphology have more types of word-forms with lower token frequency than languages with less complex morphology. Types are the set of unique word-forms in a text and tokens are all instances of word-forms. One way to measure degree of synthesis is hence type-token ratio, another one is using trigonometry in token-type diagrams [PMA09]. The simplest kind of morphology is concatenative morphology distinguishing parts of words (morphemes) of three kinds: stems (lexical element), and prefixes and suffixes (grammatical elements preceding or following stems). Given the distribution of lexical elements in parallel texts is known in one language, the forms of a lemma can be extracted for all languages with considerable accuracy. In the set of extracted forms invariant strings will be stems and variable strings prefixes and suffixes (depending on position). This allows us to measure the degree of prefixing and suffixing in different languages (as done manually in [Dry05]). The amount of case marking can be estimated effectively from extracting just the forms of proper names by the same method since proper names do not usually vary in any other grammatical category except case [Wäl12]. Finally, the amount of analytic vs. synthetic marking of negation (whether negation is expressed in a word, as in English *not* or as an affix, as in Czech *ne-*) is measured with two different algorithms [Wälth]. All extracted features are continuous.

2.2. External data sources

The online version of the Ethnologue (`http://www.ethnologue.com`) contains a comprehensive listing of the living languages of the world together with their genealogical relationships. The current version of the database contains demographic, geographic and linguistic information on 6,909 languages, also recording some already extinct languages which have gone out of use since the first edition of the database 50 years ago. We take the information given in the Ethnologue database as a standard for our genealogical hierarchies in the visualization. This information can easily be changed in the tool if a user wants to employ a different family tree for individual groups of languages.

The database of the Automated Similarity Judgment Program (ASJP) is a collection of Swadesh list items for a large number of languages. The so-called Swadesh list is an attempt to restrict the number of lexical items that have to be collected for comparing the shape of words of individual languages to a manageable basic subset of the vocabulary (typically 100 or in the case of the ASJP data 40 items) which is culturally neutral and which is expected to be relatively stable over time, i.e., it includes items which are less prone to be borrowed from other languages. Version 13 of the database [WMV*10] comprises 207,290 lexical items from 4,816 different languages and dialects and also includes areal information which is given in the form of single point longitude/latitude geo-coordinates for each language.

The World Atlas of Language Structures (WALS; [DH]) is a database of structural features of languages which have been collected by various authors on the basis of descriptive material. Currently, the database includes 76,492 data points for 2,678 languages and 198 language features, with the number of languages for individual features ranging from only 5 to 1,519. Currently, on average a feature has entries for less than 15% of all languages, i.e., the data is rather sparse. In addition, the database also comprises genealogi-

cal and areal information for each language. As in the ASJP data, the areal information is given in the form of single point longitude/latitude geo-coordinates for each language. In contrast to the automatically extracted features the structural properties in the database are either recorded as nominal (e.g., word order types such as SVO or SOV) or ordinal (e.g., the complexity of syllable structures) values. The online version of the database (`http://wals.info/`) provides world maps for each structural feature where feature values for individual languages are mapped to colored dots. However, this map display is not complemented with genealogical information.

3. Analysis Goals and Tasks

In principle, there are four reasons why languages can share a certain feature (cf. [Com89]). Beyond the trivial case (a) where all languages share the feature and (b) features are shared by chance, linguists are interested especially in whether (c) features are shared due to genealogical inheritance or (d) due to areal contact (borrowing).

In order to be able to distinguish between (c) and (d) both genealogical (hierarchical) and areal (geo-spatial) information has to be combined in one visualization. The combination of both types of information can be achieved from two different angles. On the one hand, it can be checked for a given areal pattern whether the languages that are included all belong to the same family and thus lead to a clustering of the same feature at a certain region of the world or whether there is a real contact situation where unrelated or distantly related languages share a feature. On the other hand, it is of interest to check for a given family whether the feature values are the same or similar for individual languages or whether an unusual feature value occurs, which can be attributed to the fact that the language is spoken in some other area and therefore might have borrowed that divergent feature from the languages in that area.

A further advantage of the Sunburst visualization is that a considerable number of features (both nominal and numeric) can be visualized together without much data reduction which allows for a direct introspection of the degree of homogeneity or heterogeneity of a large dataset or parts of it (some families or features being more heterogeneous than others). There are no visualizations of typological data up to now that can achieve this goal.

Hence, the visualization has multiple aims. It helps linguists to formulate hypotheses about feature inheritance or borrowing in individual cases and to assess the degree of homogeneity of the complex datasets or parts of them in direct comparison to each other. The visualization further allows to see which features are more stable genealogically than others.

4. Related Work

The core part of our approach is a visual display of the language genealogy, which is a hierarchical data structure. In the literature, several basic approaches for displaying hierarchical data can be found: Node-link tree diagrams, Icicle Plots [KL83], Treemaps [JS91], and radial space-filling layouts like the Information Slices [AH98] and later the Sunburst display [SZ00].

For our purposes we need to compare multiple language features across languages within and across different hierarchical categories. To the best of our knowledge so far no other approaches have been published that pursue this as a main goal. However, there are different approaches that plot relations among different nodes in a hierarchy, which will be described in Section 4.1. In addition, there exist approaches that combine geo-spatial information and hierarchical data, which will be described in Section 4.2.

4.1. Hierarchical and relational data

A visualization that integrates both hierarchical and relational data are ArcTrees [NSC05], a combination of a Treemap with an arc diagram. The Treemap grows only in horizontal direction and linking arcs connect two nodes of the hierarchy if they are related. Another technique with a similar purpose are Holten's Hierarchical Edge Bundles [Hol06], which can be combined with different hierarchical visualizations. Both techniques, however, require the link space to be rather sparsely populated and do not scale for fully connected graphs. A further possibility to combine hierarchical data with relational clues is to provide a matrix display that shows relations in the cells. Either the axes elements of the matrix are leaf nodes of a hierarchical node-link tree structure as in the Matrix Browser [ZKB02] or the hierarchy is conveyed through a Treemap-like recursive subdivision of the matrix as in [vH03]. Fully connected graphs are not a problem for these latter approaches, but overall only a limited number of nodes can be displayed; otherwise the matrix will grow too big. For all of the mentioned techniques, there is no intuitive way to use the visualization for feature comparison.

4.2. Hierarchical and geo-spatial data

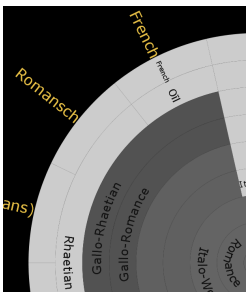
An approach that combines hierarchical and geo-spatial data are Flow Maps [PXY*05, BSV11]. Flow Maps lay a tree structure over a map in order to indicate geo-spatial movements (flows). The tree structure is the result of a hierarchical clustering of locations. Thus, the hierarchy directly depends on the geo-locations and is of a binary nature. In our scenario, in contrast, the hierarchy is predefined and not directly related to geo-spatial distributions. A different approach to combine geographic and hierarchical information into one visual display is to consider spatial ordering when creating space-filling rectangular layouts like Treemaps. One option is to take longitude and latitude values into account when splitting the Treemap rectangles as in Mansmann's Geographic HistoMap Layout [MKN*07].Wood and Dykes [WD08] follow the same fundamental idea with their

Spatially Ordered Treemaps. Later Slingsby et al. [SDW09] suggest a further version of geographically ordered space-filling rectangular layouts. All of the mentioned approaches share the property that either the upper levels of the hierarchy are geospatial, e.g., areas and subareas, or that in each leaf node geo-spatial distributions have to be displayed. The latter option causes difficulties when having many geo-spatial locations and limits the possibilities for conveying feature information in leaves.

5. Our Approach

This section contains detailed explanations about our approach which are also illustrated and exemplified in Figure 1 for a better understanding. Our goal is to give a complete overview of all available language resources integrating automatically extracted and manually edited language features with genealogical and areal information into one visual analytics system. As a core part, we suggest a novel Sunburst display that was implemented using *prefuse* [HCL05] and is able to show different types of language features, even combined at the same time:

1. The homogeneity of *distance-based features* is plotted to the inner-nodes of the Sunburst. Distance-based features may be any abstract data features. The only requirement is that their distance can be calculated according to a metric distance function. Examples are the edit-distance of Swadesh lists or geographic distances among languages. Of course, for any single or multivariate quantitative feature homogeneity can also be calculated and mapped to the inner nodes.

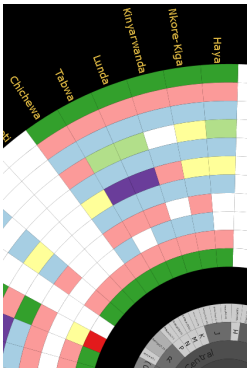


The saturation of the grey tone of an inner node, indicates whether the languages of the corresponding family on average have small distances (light grey) or large distances (dark grey). Apart from providing additional information, this coloring also helps to perceive the hierarchical relations easily.

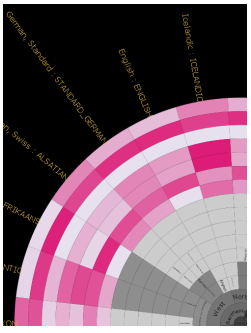
2. The *quantitative, ordinal or nominal features* are plotted to the outer rings of the Sunburst display. For each feature dimension one ring is reserved, the value for a certain language is mapped to the color, brightness or degree of fill of the ring segment belonging to that language. Examples for such values are the quantitative value showing the degree of prefixing of a language's words, as described in Section 2.1, or the nominal value of a language's word order type, as described in Section 2.2. The segments belonging to one feature dimension are aligned in one ring, readily enabling the comparison across languages in accord with the Gestalt law of continuity.

5.1. Plotting language features

Mackinlay's fundamental research [Mac86] has shown that the choice of suitable visual variables to convey information depends on the data types. In our case, the two generally most valuable visual variables, namely the x and y Position, are already used to display the hierarchy. Consequently, we pick the next best choice according to Mackinlay's research to plot the language features. This next best choice is different for quantitative, ordinal and nominal data.

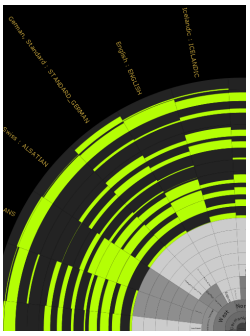


Nominal Data: For the nominal data we use different color hues to encode different categories. We follow the suggestions for color maps provided by the Color Brewer (see `http://colorbrewer2.org/`). It has to be remarked that we aim to create two notably different color maps keeping the hues as disjoint as possible. The reason is that if two adjacent nominal feature segments have the same color, they appear as a visual pattern calling the attention of the user. This is beneficial if the two segments are located within the same feature ring, i.e., two closely related languages share the same feature value. However, if the two segments are located in different feature rings they are meaningless. To avoid the second case, two adjacent nominal feature rings get different color maps that are as disjoint as possible. To do so, we use a color map from color brewer that contains 11 different colors for nominal data, which are about as many colors as can be readily distinguished. As typically a nominal feature dimension in our data has only 5 or 6 categories, usually we can split this color map into two disjoint color maps. In this case, the first ring gets the first colors of the color map and the second ring gets the further colors. Of course, in cases with more nominal categories, it cannot be guaranteed that the color maps for the two adjacent rings do not overlap, but at least the number of overlapping colors is minimized. Missing values can be colored in white or grey.



Ordinal Data: For conveying ordinal features Mackinlay identifies different density or color saturation values to be suitable. We decided to take different color saturation values. Thus, we divide the spectrum of all color saturation values by the number of different ordinal values for a feature. Like that, we get a set of ordered color tones of the same color hue that can be distinguished easily. We decided to take a pink hue as this sticks out and is

sufficiently dissimilar to the hues used for the nominal data.



Quantitative Data: For quantitative data it makes sense to use the variable *size* in order to reveal relative differences among the feature values. The quantitative feature rings in our approach show values in a histogram, where the height of the bars corresponds to the feature value. Again we chose a hue that is dissimilar to the hues used for the nominal data.

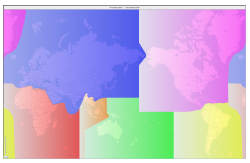
Visually conveying the data type

The meaningful mapping, however, is not the only reason why we decided to use different visual variables for the different data types. A further advantage is that the user is able to immediately recognize the data type of a random feature ring shown to him. This is especially valuable if one user creates a visualization selecting a set of features with mixed data types and shows it to another user. Yet, we still enable users to change the pre-configured mapping, for example s/he can also choose color saturation values to represent quantitative features.

5.2. Integrating areal information

Geographic information is also integrated into the Sunburst display. We have already the option to map the homogeneity (average) of geographic distances among languages in the same branch to the grey tone of the corresponding node in the hierarchy. Further options are explained in the following.

Macro areas as nominal dimension



As mentioned before, the world can be divided into macro areas, within which language contact is known to have happened a lot and across which language contact used to be rare. The Pacific Ocean lies in the middle, because historically the Atlantic Ocean has been a real diffusion border until colonialization, unlike the Pacific. We allow the user to choose between two ways of integrating the information about macro areas into our extended Sunburst visualization:

1. The macro areas can be incorporated as a new first level into the language genealogy hierarchy. That means the root of the tree has no particular meaning, next the languages are split up according to contact regions and only below according to the language genealogy.
2. The macro areas can be incorporated as the innermost ring into the display. In this case, the macro areas can be seen as another nominal data dimension and the ring segments will be colored according to the coloring of the macro areas on the world map. The user has the option to choose increasing color saturation values within macro areas either from east to west or north to south to get more detailed information about the location of a language.

Interactive linking of the world map

Of course, the mentioned possibilities only give some first hints about geography. To explore the exact geo-spatial distribution of languages the Sunburst display is interactively coupled to the world map. Through linking and brushing, the geo-spatial distribution of all languages belonging to a selected branch is displayed on this separate world map. Each language has exactly one point on the world map, because this is what the data gives us. A small circle will be displayed at the language position colored according to a user-selected language feature. At the same time, the user can select arbitrary areas on the world map and create a Sunburst containing only those languages that are located in the selected area. In addition, the user has the option to ignore the coloring of the macro areas and create a bipolar color map for the selected area as shown in Figure 4.

5.3. User Interaction

The interactive linking between the Sunburst and the world map is only one way of interacting with the display. The user is interactively involved in the data analysis process right from the start, see Figure 2. For example, s/he is asked to specify the data types of the feature dimensions and able to change it anytime, in case of errors. Both the world map and the Sunburst enable panning and zooming interaction. In the Sunburst, the user can select to focus on different aspects, e.g., single language families, languages and features that s/he is currently interested in. While this information will be highlighted all other information will be covered with a semi-transparent dark grey color tone in order not to distract her/him. In addition, the global data distribution for the highlighted feature ring will be displayed in a further panel.

6. Application Case Studies

The development of the tool was an interdisciplinary effort involving linguistic researchers right from the start. We met regularly to assure a correct understanding of the linguistic data and tasks and discuss further steps. In the following case studies we would like to report on the productive work with the tool and our first findings. In order to be able to discriminate between cases of language contact and inheritance from a proto-language (cf. Section 3) it is necessary to combine both the genealogical (hierarchical) and the areal (geo-spatial) information about languages. As mentioned before, the impact of a contact scenario can be inspected from two perspectives: (i) looking at geographical distributions (areal

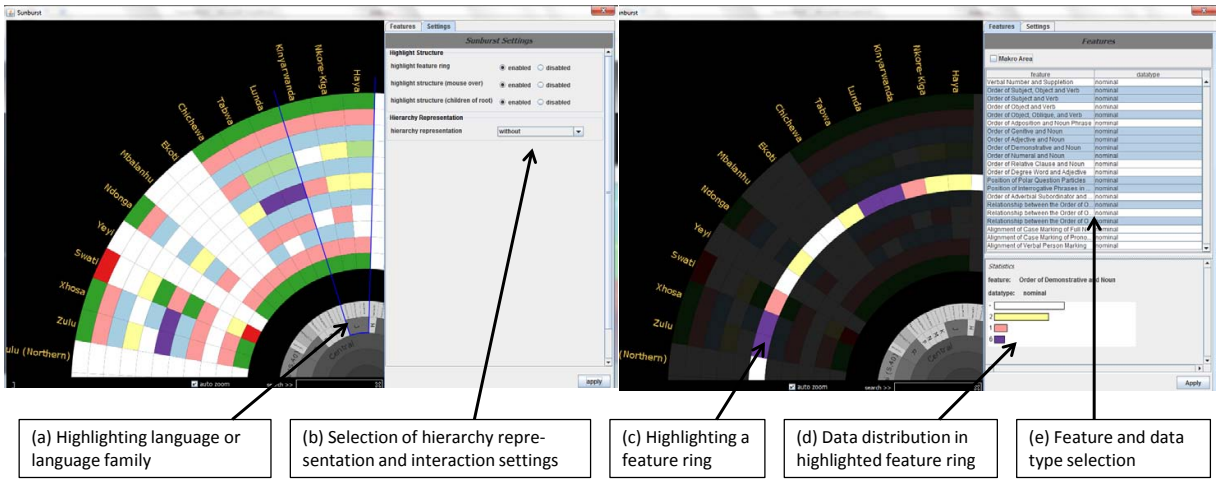


Figure 2: *User Interaction with the Sunburst and feature rings*

patterns) and checking whether all languages in the given area are from the same family; (ii) looking at a particular family (or genus) and checking whether all languages exhibit the same feature values and are spoken in the same region. We will concentrate on the latter aspect with two application case studies of our Sunburst visualization which enables the user to check for a larger amount of features whether there are outliers within the family that result from the fact that a language is spoken in a different area. As to the language properties, we experimented mainly with the automatically extracted features which have been inferred from the parallel Bible texts as expounded in Section 2.1. These features give a good approximation of what linguists have analyzed manually and are also interesting for contact situations for which the visualizations are designed. In order to test the visualization for its usability, a number of language families have been inspected by the domain experts among us. Several interesting findings could be inferred from the visual representation of the features.

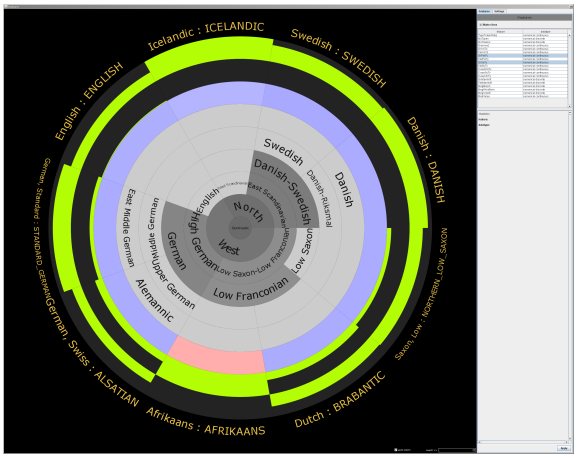


Figure 3: *Detailed look into two quantitative features for the Germanic languages.*

Case Study 1

First, we will concentrate on a particular case which can be most easily explained for non-experts. For this purpose, we look at the more familiar Indo-European language family, which also includes the prominent European languages English, French or German. Figure 1 shows the Sunburst representation of the Indo-European languages in our sample and their hierarchical structure of subfamilies (genera). In addition, the innermost ring of the visualization shows the color-coded macro-area in which the respective language is spoken. It can be seen at-a-glance that the languages are spoken in the same macro-area (Eurasia), with the sole exception of Afrikaans, which is located on the African continent. Furthermore, Afrikaans can easily be detected as an outlier with respect to its neighboring languages, i.e. many feature values deviate strongly. Part of this effect is due to the fact that some features are correlated, which can be seen when looking at their distribution over all languages, however, Afrikaans is visually salient independent from that. For the sake of simplicity, we select only the family of Germanic languages and look at just two features in the Sunburst visualization, namely the synthesis parameters of prefixation (morphological material occurring before the stem) or suffixation (morphological material occurring after the stem). In the visualization in Figure 3, both features are depicted in the outer rings of the Sunburst, with the prefixation feature as the inner ring of the two and suffixation as the outermost ring. When looking at both

feature rings, it can immediately be seen that Afrikaans is not only peculiar because of its areal status but also regarding the feature values that it has. In comparison to the adjacent (West) Germanic languages Afrikaans has a higher prefixation and a lower suffixation value. This is particularly interesting because it is in a contact situation with surrounding African languages (our sample contains the Bantu languages Zulu and Xhosa, which are also spoken in South Africa). Bantu languages are notorious for their extensive use of prefixes to convey grammatical meaning on the verb. The comparatively higher prefixation value for Afrikaans thus might be caused by the influence of the Bantu morphological patterns. On closer inspection, however, it turns out that Afrikaans makes extensive use of the perfect construction involving the past participle with *ge-* (similar to Dutch or German). The synthetic past tense forms (the so-called *imperfect tense*) where a further distinction for different persons (first person singular, third person singular, etc.) is made in suffixes have disappeared except for a few vestigial cases [Don93]. The fact that a further distinction in suffixes does not exist with the past participles, which are now dominant in the language to convey reference to a past event, results in a lower suffixation value for the language with respect to other (West) Germanic languages. Whether the use of the perfect instead of the past tense is a direct influence of the contact languages or merely due to the geographic separation of Afrikaans with respect to other Germanic languages (especially its sister language Dutch), however, remains to be investigated. Yet the visualization easily enables the linguist to check for such suspicious patterns which can later be inspected in more detail.

Case Study 2

While the Germanic Languages in general are well-studied, for other language families the available knowledge can be very limited. For only a few of the numerous languages spoken in Papua New Guinea grammar books are available. Translations of Bible texts, however, can be gathered for quite a lot of them (`http://www.pngscriptures.org`). The features automatically extracted from those can be seen in Figure 4. The following findings could be obtained by the domain experts among us, without having closer knowledge about the individual languages: Austronesian languages are rather more homogeneous in their feature values than Papuan languages which is in line with their well established genealogic relationship. Another quite homogeneous group is Huon-Finisterre (high synthesis, no morphological negation, very little prefixing, much suffixes among which case suffixes). Within East New Guinea, genealogical subgroupings clearly emerge when several features are considered. The Eastern subgroup, for instance, is characterized by high synthesis, analytic negation, no prefixes, moderately high suffixing and case suffixes. The Chimbu subgroup is distinguished by morphological negation, lack of case and rather low degree of synthesis. There is much heterogeneity in East Papuan, well in line with the fact that this is not quite an established family such as Austronesian. Within the Austronesian family the subfamily of the Papuan Tip languages can be distinguished both with respect to certain features and the geo-locations.

7. Discussion

In this section we would like to discuss different aspects, problems, and open issues of this project as well as lessons learned from our interdisciplinary collaboration.

For the application development the involvement of domain experts from the very beginning on was extremely useful, much more than expected. The incorporation of a deep understanding of the domain and its data prevented us to go into a wrong direction when designing the application. The concrete analysis tasks of the domain experts provided a good guidance to design the tool.

A problem from the analysis perspective is the sparseness of the data. While the genealogical information for about 6,900 languages is available, most features are only available for a few hundred languages. However, linguists are about to collect more and more such data and for the future it is to be expected that the data issue will become less critical.

The number of languages is much larger than the number of different features. Currently, we have less than 250 features. The number of feature rings, consequently, is limited and the whole data set can still be visualized on a large high-resolution display. However, given that the number of features to be displayed grows heavily in the future, an Icicle plot may be more suitable than a Sunburst when displaying the whole dataset at once. The reason is that each additional feature ring for the Sunburst needs more space than the previous one and the display grows in x and y direction. An Icicle plot would still allow to map information to the inner nodes of the hierarchy and the display would only grow in y direction and become more quadratic as the number of features approaches the numbers of languages. However, we could observe that in a typical analysis case only a limited set of features is available or of interest. For this setting the Sunburst makes a better use of the screen space.

While the hierarchically structured genealogical information is the core of our display, the integration of geo-spatial views yet is on a fairly basic level. Several open issues can be identified that could help to improve the integration in the future. First, working with distorted maps might be useful to grant more space to data-wise densely populated regions. Ideally, the distortion would be constantly re-calculated according to the current selection of languages, but this does not work in interactive time with current cartogram algorithms. Secondly, the division of the world map into macro areas is linguistically motivated and easy to understand, but coarse grained. For arbitrary smaller regions the user can create customized color maps, similar to a two dimensional color mapping of location proposed by Wood and Dykes [WD08]. Fur-

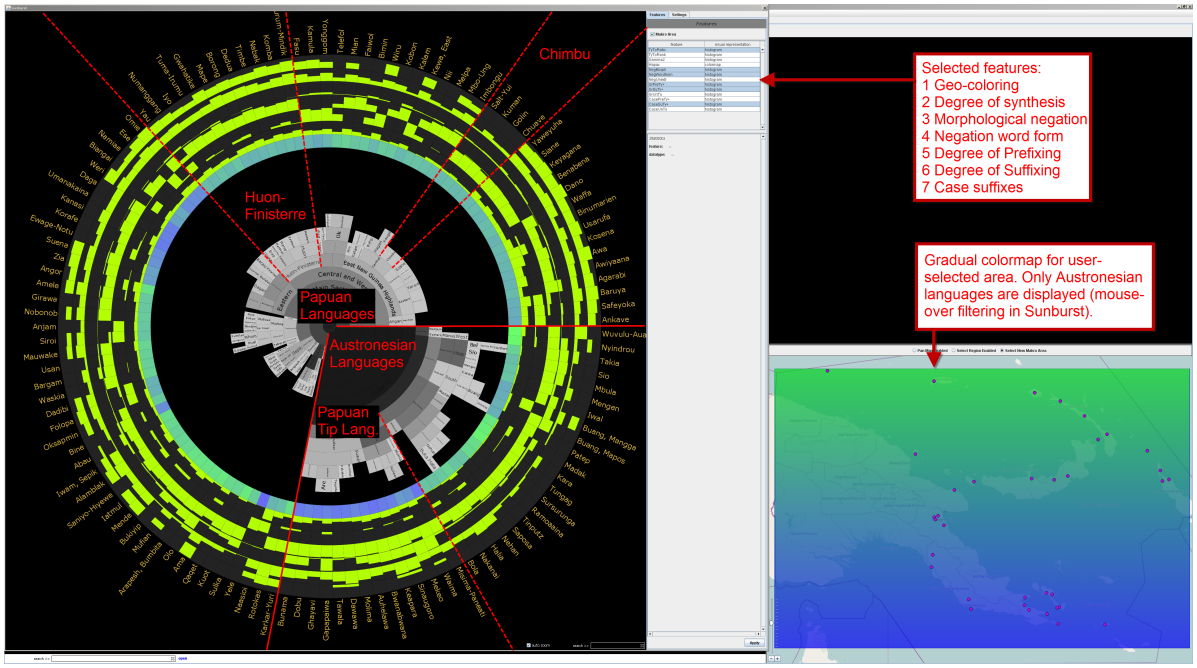


Figure 4: *High-resolution screenshot showing automatically extracted features for languages from Papua New Guinea.*

ther, it is known that populations, and with them languages, are more likely to spread within the same climate zone than across climate zones [Gülss]. The actual likelihood of language spread in each direction also depends on natural borders like seas, mountains and deserts. It would certainly be an interesting topic for future interdisciplinary work to generate a color map that encodes a "spread"-distance between languages.

8. Conclusion

In this paper we introduced a new field of application for Visual Analytics: Historical Comparative Linguistics, and Linguistic and Areal Typology. We provided background information about the research in this field including concrete tasks and requirements and available data sources. In our approach we demonstrated how linguistics research can profit from Visual Analytics. In particular, we suggested an extended Sunburst visualization with *feature rings* in order to enable the comparison of several features at once in the context of language genealogy. We discussed different ways to design the feature-rings that are optimized for the data types nominal, ordinal and quantitative. In a second step, we linked the hierarchical display with a geo-spatial visualization and suggested ways to integrate the geo-spatial information into our Sunburst.
Domain experts were involved into the development from the beginning on to assure that their tasks and data were correctly understood. Their suggestions were considered during the development. In the end, they used the final version of our tool and were able to generate new hypotheses relevant to their field and confirm old ones. Visualization also showed to be a good means to discuss hypotheses and theories.
While the domain for which this application was designed seems to be rather narrow, there actually is a whole research community interested in the outlined analysis tasks. One of the data resources investigated, the WALS data, has even become a standard resource for teaching in Linguistics. In addition, further research communities with related tasks, like the variation genetics field in biology, could potentially profit from the presented application.
In our future work we aim to extend the geo-spatial component of the approach and experiment with further interaction techniques.

Acknowledgment

We are grateful to Östen Dahl for help with bringing some of the N.T. data into an easily processable form and to Ljuba Veselinova for help with the language data. We would also like to thank Michael Cysouw and Miriam Butt for valuable suggestions and comments. This work was partially funded by the Research Initiative for "Computational Analysis of Linguistic Development" (CALD) at the University of Konstanz, Germany.

References

[AH98] ANDREWS K., HEIDEGGER H.: Information slices : Visualising and exploring large hierarchies using cascading , semi-circular discs. *Information Visualization* (1998), 9–12. 4

[BSV11] BUCHIN K., SPECKMANN B., VERBEEK K.: Flow map layout via spiral trees. *IEEE Trans. Vis. Comput. Graph. 17*, 12 (2011), 2536–2544. 4

[Com89] COMRIE B.: *Language Universals and Linguistic Typology*. Oxford: Basil Blackwell, 1989. 4

[CW07] CYSOUW M., WÄLCHLI B.: Parallel texts: using translational equivalents in linguistic typology. *Sprachtypologie und Universalienforschung STUF 60*, 2 (2007), 95–99. 2

[DH] DRYER M., HASPELMATH M.: The World Atlas of Language Structures Online. Munich: Max Planck Digital Library. http://wals.info/. 1, 3

[Don93] DONALDSON B. C.: *A Grammar of Afrikaans*. Berlin: Mouton de Gruyter, 1993. 8

[Dry92] DRYER M. S.: The Greenbergian word order correlation. *Language 68*, 1 (1992), 80–138. 1

[Dry05] DRYER M. S.: Prefixing vs. suffixing in inflectional morphology. In *The World Atlas of Language Structures*, Haspelmath M., Dryer M. S., Gil D., Comrie B., (Eds.). Oxford: Oxford University Press, 2005, ch. 26. 3

[Gre60] GREENBERG J. H.: A quantitative approach to the morphological typology of languages. *International Journal of American Linguistics 26* (1960), 178–194. First published in Spencer, Robert F. 1954. Fs. for Wilson D. Wallis. Method and perspective in anthropology. University of Minnesota Press. 3

[Gülss] GÜLDEMANN T.: "Sprachraum" and geography. In *The Handbook of Language Mapping*, Handbooks of Linguistics and Communication Science. Berlin: Mouton de Gruyter, in press. 9

[HCL05] HEER J., CARD S. K., LANDAY J. A.: prefuse: a toolkit for interactive information visualization. In *Proceedings of the 2005 Conference on Human Factors in Computing Systems, CHI 2005, Portland, Oregon, USA, April 2-7, 2005* (2005), van der Veer G. C., Gale C., (Eds.), ACM, pp. 421–430. 5

[Hol06] HOLTEN D.: Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Trans. Vis. Comput. Graph. 12*, 5 (2006), 741–748. 4

[JS91] JOHNSON B., SHNEIDERMAN B.: Tree maps: A space-filling approach to the visualization of hierarchical information structures. In *IEEE Visualization* (1991), pp. 284–291. 4

[Juo08] JUOLA P.: Assessing linguistic complexity. In *Language Complexity. Typology, Contact, Change*, Miestamo M., Sinnemäki K., Karlsson F., (Eds.). Amsterdam: Benjamins, 2008, pp. 89–108. 3

[KL83] KRUSKAL J. B., LANDWEHR J. M.: Icicle Plots: Better Displays for Hierarchical Clustering. *The American Statistician 37*, 2 (1983), 162–168. 4

[Mac86] MACKINLAY J. D.: Automating the design of graphical presentations of relational information. *ACM Trans. Graph. 5*, 2 (1986), 110–141. 5

[MKN*07] MANSMANN F., KEIM D. A., NORTH S. C., REXROAD B., SHELEHEDA D.: Visual analysis of network traffic for resource planning, interactive monitoring, and interpretation of security threats. *IEEE Trans. Vis. Comput. Graph. 13*, 6 (2007), 1105–1112. 4

[Nic92] NICHOLS J.: *Linguistic Diversity in Space and Time*. Chicago: The University of Chicago Press, 1992. 1

[NSC05] NEUMANN P., SCHLECHTWEG S., CARPENDALE M. S. T.: Arctrees: Visualizing relations in hierarchical data. In *EuroVis05: Joint Eurographics - IEEE VGTC Symposium on Visualization, Leeds, United Kingdom, 1-3 June 2005* (2005), Brodlie K., Duke D. J., Joy K. I., (Eds.), Eurographics Association, pp. 53–60. 4

[PMA09] POPESCU I.-I., MAČUTEK J., ALTMANN G.: *Aspects of Word Frequencies*. Lüdenscheidt: RAM, 2009. 3

[PXY*05] PHAN D., XIAO L., YEH R. B., HANRAHAN P., WINOGRAD T.: Flow map layout. In *IEEE Symposium on Information Visualization (InfoVis 2005), 23-25 October 2005, Minneapolis, MN, USA* (2005), IEEE Computer Society, p. 29. 4

[RB98] RIJKHOFF J., BAKKER D.: Language sampling. *Linguistic Typology 2*, 3 (1998), 263–314. 2

[RMB*10] ROHRDANTZ C., MAYER T., BUTT M., PLANK F., KEIM D. A.: Comparative visual analysis of cross-linguistic features. In *Proceedings of the International Symposium on Visual Analytics Science and Technology (EuroVAST 2010)* (2010), Kohlhammer J., Keim D. A., (Eds.), pp. 27–32. 2

[SDW09] SLINGSBY A., DYKES J., WOOD J.: Configuring hierarchical layouts to address research questions. *IEEE Trans. Vis. Comput. Graph. 15*, 6 (2009), 977–984. 5

[SZ00] STASKO J. T., ZHANG E.: Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *INFOVIS* (2000), pp. 57–. 4

[TFES11] THERÓN R., FONTANILLO L., ESTEBAN A., SEGUÍN C.: Visual analytics: A novel approach in corpus linguistics and the nuevo diccionario histórico del español. *III Congreso Internacional de Lingüística de Corpus* (2011). 2

[vH03] VAN HAM F.: Using multilevel call matrices in large software projects. In *9th IEEE Symposium on Information Visualization (InfoVis 2003), 20-21 October 2003, Seattle, WA, USA* (2003), IEEE Computer Society. 4

[Wäl12] WÄLCHLI B.: Indirect measurement in morphological typology. In *Methods in Contemporary Linguistics*, Ender A., Leemann A., Wälchli B., (Eds.). Berlin: Mouton de Gruyter, 2012. 3

[Wälth] WÄLCHLI B.: Algorithmic typology, aggregating without features and going from known to similar unknown categories within and across languages. In *Aggregating Dialectology and Typology*, Szmrecsanyi B., Wälchli B., (Eds.). Berlin: Mouton de Gruyter, Forth. 3

[WD08] WOOD J., DYKES J.: Spatially ordered treemaps. *IEEE Trans. Vis. Comput. Graph. 14*, 6 (2008), 1348–1355. 4, 8

[WMV*10] WICHMANN S., MÜLLER A., VELUPILLAI V., BROWN C. H., HOLMAN E. W., BROWN P., SAUPPE S., BELYAEV O., URBAN M., MOLOCHIEVA Z., WETT A., BAKKER D., LIST J.-M., EGOROV D., MAILHAMMER R., BECK D., GEYER H.: The ASJP database (version 13). URL: http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm, 2010. 3

[Yps09] YPSILANTI: A digital library of language relationships. Ypsilanti, MI: Institute for Language Information and Technology (LINGUIST List), Eastern Michigan University. http://multitree.org/, 2009. 2

[ZKB02] ZIEGLER J., KUNZ C., BOTSCH V.: Matrix browser: visualizing and exploring large networked information spaces. In *CHI Extended Abstracts* (2002), Terveen L. G., Wixon D. R., (Eds.), ACM, pp. 602–603. 4