Language Documentation & Conservation Special Publication No. 3 (August 2012):
Potentials of Language Documentation: Methods, Analyses, and Utilization,
ed. by Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann,
Dagmar Jung, Anna Margetts, and Paul Trilsbeek pp. 96–104
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4522

Visualization and online presentation of linguistic data

Hans-Jörg Bibiko

Max Planck Institute for Evolutionary Anthropology, Leipzig

This contribution gives an introduction to state-of-the-art techniques for the visualization and online presentation of linguistic data and world-wide linguistic diversity, such as linguistic maps and online dictionaries, using a software environment called R. The aim is to draw linguists' attention to the possibilities offered by these techniques and to give some practical hints as to how they can be used specifically for linguistic and language documentation data.

1. R AS A TOOL FOR CREATING THE VISUALIZATION AND ONLINE PRESENTATION OF LINGUISTIC DATA. Visualization by way of diagrams, charts, animations, maps, etc. and their online presentation is an important means of enhancing the usability of linguistic data for various scientific and educational purposes and of presenting linguistic facts to the general public. Nowadays these visualizations can be created relatively easily with R, an open-source scripting-language based software environment for doing statistics and generating high-quality graphics (see www.R-project.org) . Since R is open source, a huge amount of so-called packages, i.e. libraries of functions for a particular purpose created by the world-wide community of R users, are available that extend the functionality of R enormously. A particularly important point in this context is that with R it is possible to generate so-called vector PDF graphics; these have various advantages over pixel images and can be edited and used in publications or presentations. In the following sections, I give examples of how R can be used to visualize complex data sets and their geographic distribution (Section 2), how custom maps can be generated (Section 3), and how structured linguistic data such as wordlists can be transformed into user-friendly resources such as online thematic dictionaries (Section 4).

2. GENERATING MAPS WITH R.

2.1. DISPLAYING WORDLISTS BY USING MAPS. An initial example of the application of R is illustrated by the geographic distribution of the word for "three" in about 3,000 languages. The data used in this example are from Eugene Chan's compilation of "Numeral Systems of the World's Languages" (http://lingweb.eva.mpg.de/numeral). A static wordlist from 3,000 languages is difficult to work with. Mapping this as online information helps one to visualize relationships, e.g. between cognate terms or areally diffused terms. Clicking the link in the caption of Figure 1 will display a movie illustrating such a visualization. The HTML/KML file that underlies this visualization was generated by means of R, since R is also a powerful scripting language, making it unnecessary for the user to have to acquire knowledge in classic scripting languages such as Python and Perl, etc.

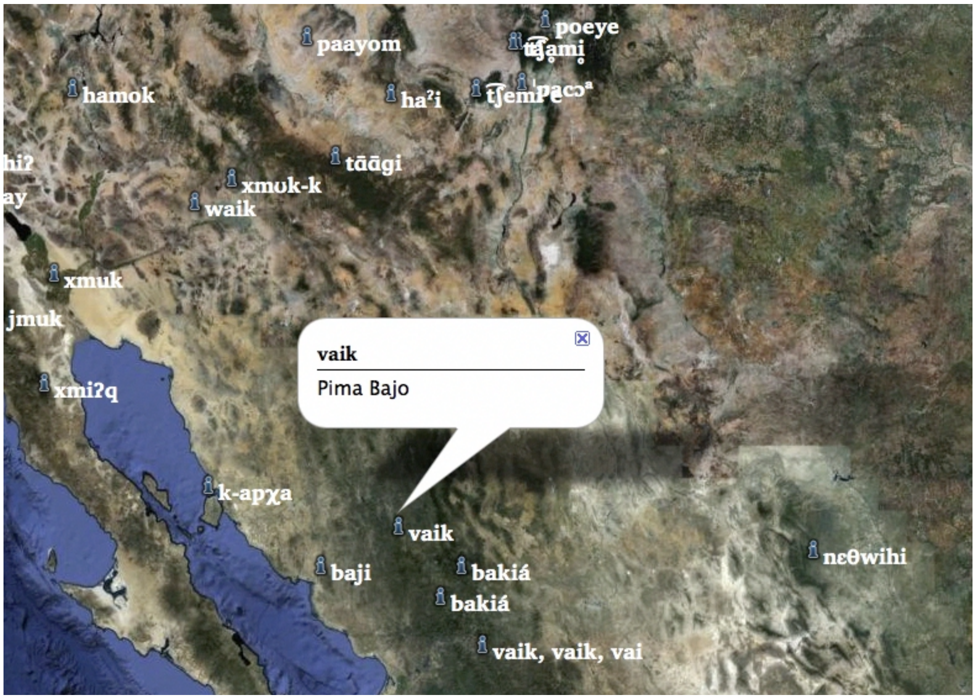


FIGURE 1: The geographic distribution of words for “three” in Northwest Mexico [video: http://youtu.be/4bb0y7lWsqQ, http://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/4522/13-wordlist.m4v]

2.2. DISPLAYING STRUCTURAL FEATURES ON MAPS. It is interesting to study the geographic distribution not only of lexical information, as in the previous example, but also of structural linguistic information, as in the World Atlas of Language Structures (WALS) (Dryer & Haspelmath 2011). Maps similar to those of WALS can be created using R, as in Figure 2; this map shows the distribution of the WALS feature “Order of Adposition and Noun Phrase” for languages spoken in China and Mongolia.

2.3. DISPLAYING VARIOUS FEATURES AND VALUES IN PIE CHARTS. In comparative and quantitative linguistics, the assignment of only one value for a language feature is often difficult. Therefore it is sometimes more meaningful to give information on the degree to which various values are true for a given feature. R can be used to create pie charts displaying this information, adding another dimension to the representation of the

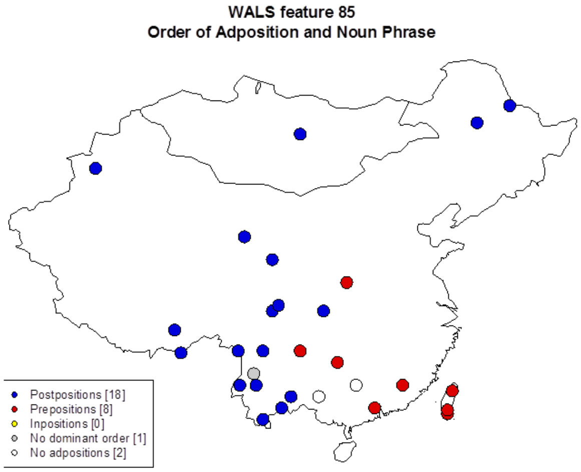


FIGURE 2: Map showing "Order of Adposition and Noun Phrase" for languages spoken in China and Mongolia

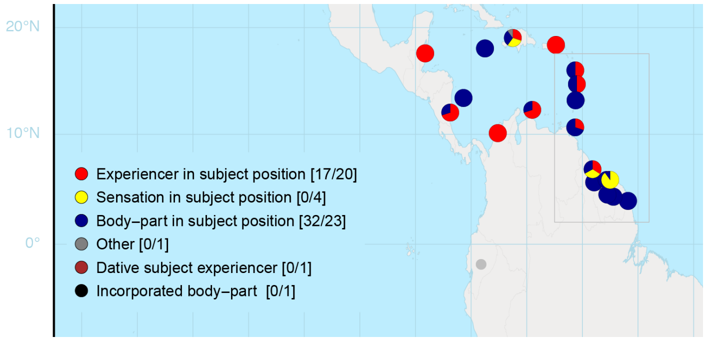

20°N
10°N
0°
● Experiencer in subject position [17/20]
● Sensation in subject position [0/4]
● Body–part in subject position [32/23]
● Other [0/1]
● Dative subject experiencer [0/1]
● Incorporated body–part [0/1]

FIGURE 3: Values for "Experiencer constructions: 'headache' " from APiCS

geographic distribution of linguistic features. For example, Figure 3 shows the values for a feature called "Experiencer constructions: 'headache' " from *The Atlas of Pidgin and Creole Language Structures* (APiCS) (Michaelis et al. forthcoming) in various languages of the Caribbean area.

Figure 4, taken from "The electronic World Atlas of Varieties of English" (eWAVE) (Kortmann & Lunkenheimer 2011), shows how different features can be represented in pie charts. The color of the bottom right section of the pie charts represents the values of the feature "She/her used for inanimate referents", the top section of the feature "Generalized third person singular pronoun: object pronouns", and the bottom left section of the feature "Me instead of I in coordinate subjects". Such a representation allows one to visually inspect if and where such features correlate.

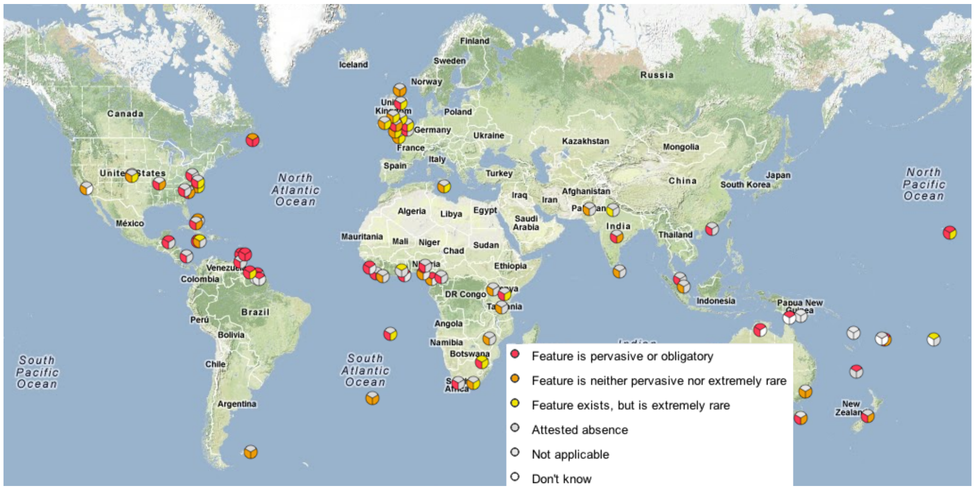


FIGURE 4: Simultaneous display of three features in eWAVE

3. INTERACTION BETWEEN R AND GEOGRAPHIC INFORMATION SYSTEMS (GIS).

3.1. USING GIS DATA TO CREATE HIGH-RESOLUTION MAPS. The contextualization of language documentation data often requires detailed custom maps of the often remote areas where the documented languages or dialects are spoken. Such maps can be created by freely available datasets stored as part of so-called geographic information systems (GIS) (see http://wiki.gis.com/wiki/index.php/Geographic_information_system). R is able to use the freely available GIS data with the packages "maptools" and "sp", as well as others. This is especially useful for generating maps of smaller scale, e.g. a smaller island.

Each custom map requires a particular resolution. For instance, a very rough resolution of the Hawaiian island O'ahu, as provided by one of the various packages developed within R itself, may be appropriate for a map of the world or a larger region (Figures 5a–5c); but the exact locations of small language or dialect communities often require a much higher resolution, and this can only be provided by GIS data sets, which can be imported into R (Figure 5d). With these data sets it is possible to zoom in to very fine levels of granularity (Figure 5e).

3.2. ADDING FURTHER INFORMATION TO MAPS. There are many data sets available containing further potentially relevant information for creating language maps, including population data, climate zones, land use, etc. For instance, topographic information, such as elevation, is also freely available and can be added to maps created in this way. For example, the GIS topographic data set "Oahu, HI 1/3 arc-second MHW DEM" (http://www.ngdc.noaa.gov/dem/squareCellGrid/download/3410) can be added as a background to the map of O'ahu (Figure 6).

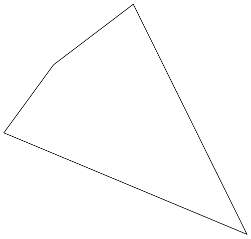


FIGURE 5a: Map created with the R package “maps” – 5 points

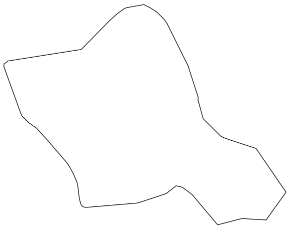


FIGURE 5b: Map created with the R package “sp” (wrld_simple data set) – 50 points

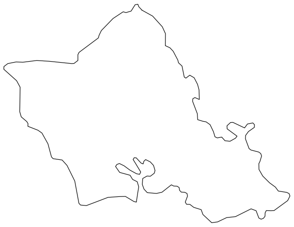


FIGURE 5c: Map created with the R package “mapdata” – 173 points

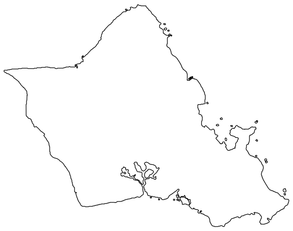


FIGURE 5d: Map created with the import of the GIS data set “gshhs” [2] – 2254 points

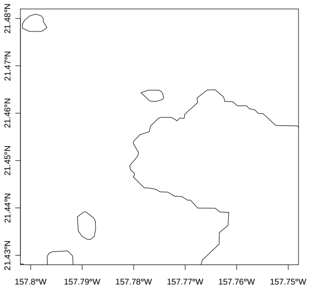


FIGURE 5e: High-resolution detail of map created with the GIS data set "gshhs"

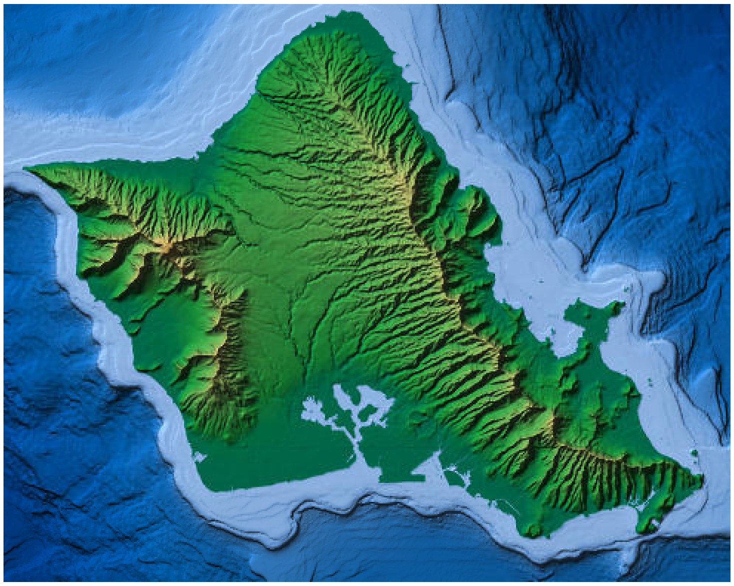


FIGURE 6: Map of O'ahu using the GIS data set Oahu, HI 1/3 arc-second MHW DEM [http://www.ngdc.noaa.gov/dem/squareCellGrid/download/3410]

3.3. DRAWING GEOREFERENCED MAPS.

Linguists often need to mark specific areas on a map, e.g. a language location, isoglosses, etc.; in other words, they need to draw a georeferenced area, which is then compatible with the maps and further geographic information discussed so far. This can be done using R's ability to import KML files, which are, for instance, used in Google maps and Google Earth. A relatively easy way to create a self-defined georeferenced polygon, like an isogloss, and include it in a map is to use the free version of Google Earth (http://www.google.com/earth). Google Earth allows one to create and to edit a georeferenced polygon, which can be saved as a KML file and imported into R. Clicking the link in the caption of Figure 7 will display a movie illustrating this workflow.

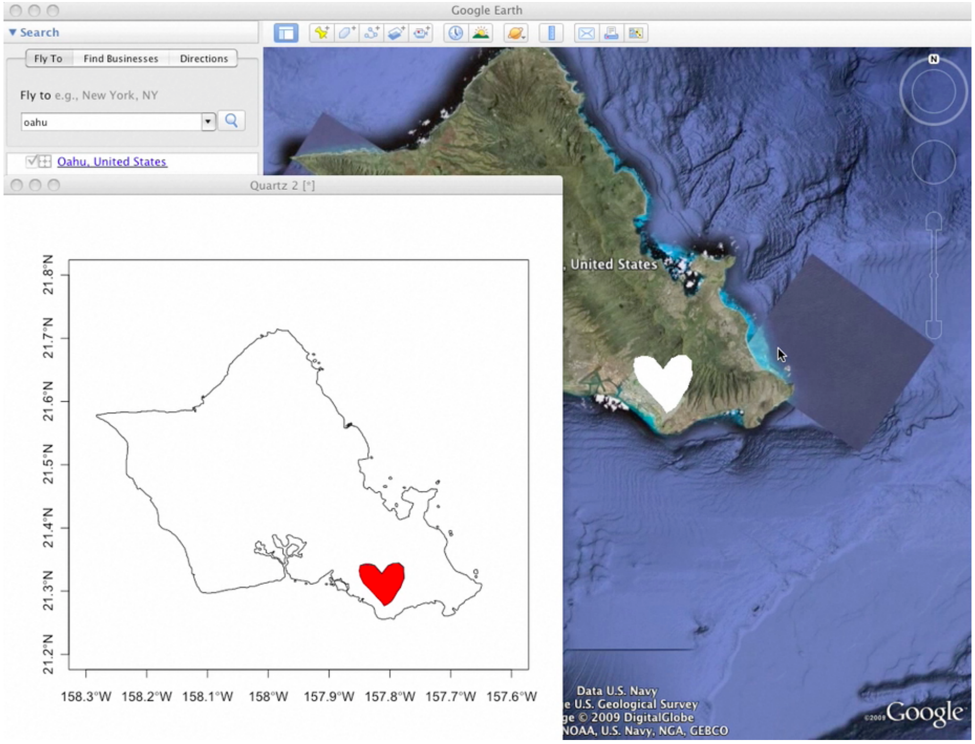
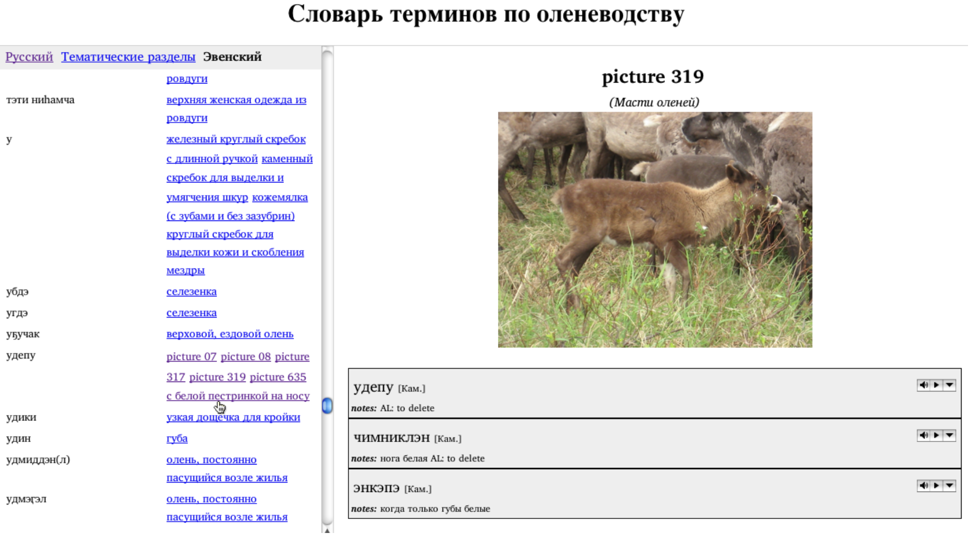


FIGURE 7: A workflow for drawing georeferenced maps
[video: http://youtu.be/mMSaSaAxCp8, http://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/4522/13-polygon.m4v]

4. GENERATING WEBSITES FROM STRUCTURED LINGUISTIC DATA. For the general public, as well as for some scientific purposes, it is useful to present linguistic data, such as thematic wordlists or entire lexica, in easily accessible formats, such as websites. If data such as wordlists are stored in formats such as Toolbox (see www.sil.org/computing/toolbox) and ELAN (see http://www.lat-mpi.eu/tools/elan), the creation of such HTML-based websites can be carried out relatively easily using R. As already mentioned, R is a powerful scripting language which works with data stored in many different formats including XML (for instance ELAN's eaf files) and plain text (for instance Toolbox files). In other words, instead of learning another scripting language like Python or Perl, such conversions (including the usage of regular expressions) can be done using R. Figure 8 shows an interactive HTML-based Even-Russian online dictionary of reindeer terms with pictures and linked sound files that was created with the help of R, using a Toolbox file as a source. The data for this example comes from Aralova et al. (in preparation).

5. CONCLUSION. The previous sections provided some examples of how R, an open-source software environment, can be used to process linguistic data, including data that is



Словарь терминов по оленеводству

Русский Тематические разделы **Эвенский**

	ровдуги
тэти ниһамча	верхняя женская одежда из ровдуги
у	железный круглый скребок с длинной ручкой каменный скребок для выделки и умягчения шкур кожемялка (с зубами и без зазубрин) круглый скребок для выделки кожи и скобления мездры
убдэ	селезенка
угдэ	селезенка
уӡучак	верховой, ездовой олень
удепу	picture 07 picture 08 picture 317 picture 319 picture 635 с белой пестринкой на носу
удики	узкая дощечка для кройки
удин	губа
удмиддэн(л)	олень, постоянно пасущийся возле жилья
удмэӷэл	олень, постоянно пасущийся возле жилья

picture 319

(Масти оленей)

удепу [Кам.]

notes: AL: to delete

ЧИМНИКЛЭН [Кам.]

notes: нога белая AL: to delete

ЭНКЭПЭ [Кам.]

notes: когда только губы белые

FIGURE 8: An online dictionary created from a Toolbox file using R

generated in language documentation, for various forms of visualizations and online presentations, which are appealing both for scientific research as well as for use by speech communities and the general public. While it is true that applying these techniques in R requires some familiarity with scripting languages in general, it is hoped that the examples given here will stimulate linguists to engage in these techniques themselves or else to cooperate with information scientists who are familiar with these environments.

REFERENCES

Amante, Christopher & Barry W Eakins. 2009. *ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis*. NOAA Technical Memorandum NESDIS NGDC-24. http://www.ngdc.noaa.gov/mgg/global/global.html.

Aralova, Natalia, Alexandra Lavrillier, Dejan Matić, Brigitte Pakendorf, Evgeniya Zhivotova & Luise Zippel. in preparation. Even dialectal dictionary of reindeer herding terminology. Leipzig: Max Planck Research Group on Comparative Population Linguistics.

Bivand, Roger S., Edzer J. Pebesma & Virgilio Gómez-Rubio. 2008. *Applied Spatial Data Analysis with R* Use R! New York: Springer.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2011. *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library. http://wals.info/.

Kortmann, Bernd & Kerstin Lunkenheimer (eds.). 2011. *The electronic World Atlas of Varieties of English (eWAVE)*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://www.ewave-atlas.org/.

Michaelis, Susanne, Philippe Maurer, Martin Haspelmath & Magnus (eds.) Huber. forthcoming. The Atlas of Pidgin and Creole Language Structures (APiCS). http://lingweb.eva.mpg.de/apics/index.php/The_Atlas_of_Pidgin_and_Creole_Language_Structures_%28APiCS%29.

National Geophysical Data Center (NGDC). A Global Self-consistent, Hierarchical, High-resolution Shoreline Database (GSHHS). http://www.soest.hawaii.edu/wessel/gshhs/.

R Development Core Team. 2011. *A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Hans-Jörg Bibiko
bibiko@eva.mpg.de