Language-Sites: Accessing and Presenting Language Resources via Geographic Information Systems

Dieter Van Uytvanck, Alex Dukers, Jacquelijn Ringersma, Paul Trilsbeek
Max Planck Institute for Psycholinguistics
P.O. Box 310, 6500 AH Nijmegen, the Netherlands

E-mail: {Dieter.VanUytvanck, Alex.Dukers, Jacquelijn.Ringersma, Paul.Trilsbeek}@mpi.nl

Abstract

The emerging area of Geographic Information Systems (GIS) has proven to add an interesting dimension to many research projects. Within the language-sites initiative we have brought together a broad range of links to digital language corpora and resources. Via Google Earth's visually appealing 3D-interface users can spin the globe, zoom into an area they are interested in and access directly the relevant language resources. This paper focuses on several ways of relating the map and the online data (lexica, annotations, multimedia recordings, etc.). Furthermore, we discuss some of the implementation choices that have been made, including future challenges. In addition, we show how scholars (both linguists and anthropologists) are using GIS tools to fulfill their specific research needs by making use of practical examples. This illustrates how both scientists and the general public can benefit from geography-based access to digital language data.

Geographic Information Systems are becoming very popular in searching language resources, presenting language related material, and merging linguistic evidence together with information from other disciplines. At the MPI for Psycholinguistics and within the DOBES project (Wittenburg, Mosel, & Dwyer, 2002) an increasing number of researchers is making use of this technique. Until now the following case scenarios were analyzed and implemented:

- offering a geographic browsing option to locate language resources as an alternative for the normal metadata-based access
- offering access to web applications to present complex resource bundles such as annotated media recordings or lexica with multimedia extensions
- organizing the data of fieldworkers according to a procedure taken at the field site that is motivated by microvariations in a language
- visualizing ethnologically relevant data via geographic markers

Currently, Google Earth[1] (Goodchild, 2006) is chosen as the primary system for rendering geographic information since it is the most popular one and many people are acquainted with it. However, the overlays that contain the linguistic and anthropological information are created as XML-based KML[2] files which means that we could easily transform them to fit with other (possibly web-based) geographic presentation frameworks. This is considered to be very important, since the availability of research information should not be restricted by the use of proprietary formats.

1. Reference to Catalogues

Many potential users of language resources do not like to look for interesting data via the formal metadata descriptions – be it in the form of electronic IMDI (Broeder, Offenga, Willems, & Wittenburg, 2001) or OLAC (Bird & Simons, 2001) catalogues. Mostly the semantics of the vocabularies are not known and the simple search options result in an information overload. The geographic metaphor is much more intuitive, since in general we relate cultures and languages with geographical areas. In addition, the geographical metaphor itself attracts much more non-experts, who are just curious. Therefore a worldwide initiative has been started by the DELAMAN network[3] of endangered language and music archives called Language-Sites[4].

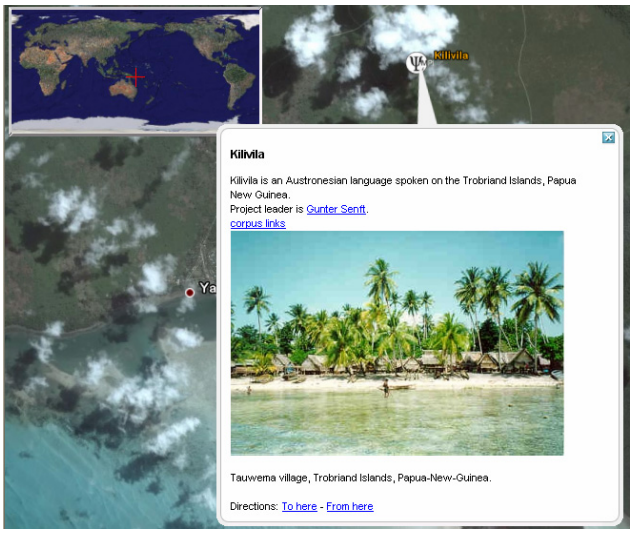

Figure 1: an example of a landmark, including a description, a picture, and a link to a relevant metadata catalogue. Data for the Kilivila language (Trobriand Islands), provided by Senft (1986).

Each archive is asked to provide landmarks for the languages of which there are resources stored in the

[1] http://earth.google.com
[2] http://code.google.com/apis/kml/
[3] http://www.delaman.org
[4] http://www.languagesites.org

archive, accompanied by a reference to the catalogue entry. Together with some introductory information an overlay entry is created in the KML format and then in turn integrated into the Language-Sites overlay. Users can explore these landmarks and go to the catalogue entry that covers all material, which is related to the selected language.

Language-Sites is an open initiative. Everyone else storing language resources and having a clear repository organization is welcome to participate. The chosen technique allows to be as fine-grained as required, i.e. it could point to a certain small region in which a certain dialect has been recorded.

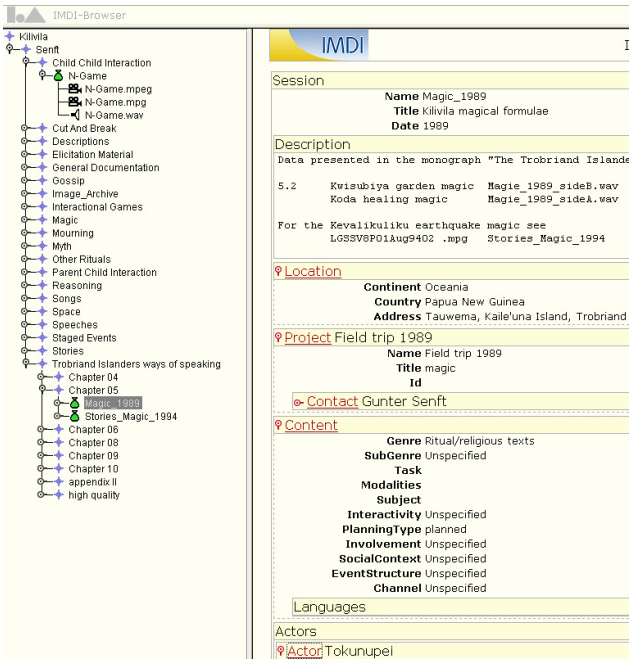

Figure 2: the view on the metadata associated to a specific placemark. This is displayed after clicking on the corpus links in figure 1.

For the moment we do not systematically include absolute geographic coordinates in our metadata catalogue yet. Doing so in the future would however make it very easy to generate landmarks with the associated information and references into the Language-Sites file.

Moreover, we have chosen not to use area descriptions to indicate a language, but singular points (i.e. a single coordinate) instead. The reason is that areas can be politically very sensitive and often heavily disputed. Simple landmarks can either refer to the place where the recording was made or as a typical place representing the geographic area where the language is or was spoken.

2. References to Complex Resource Bundles

A second application has been tested as well. Curious users, in particular those belonging to the young generation, want to access the material for a certain language they have found on a geographic map immediately. This way of exploring the metadata catalogue could provide a more appealing alternative for the traditional browsing through an abstract hierarchy. The text balloon that pops up when the user clicks on a placemark can be enriched with a descriptive text, images, sound files and video files. In addition, hyperlinks can be included in order to connect the depicted places on the map with associated web pages.

The most powerful use of this mapping system however lies in combining it with annotated movie and sound recordings or lexica with multimedia extensions.

As a first step we made it possible to invoke ANNEX[5] (Berck, Russel, Kemps-Snijders, & Wittenburg, 2005) to access annotated video or sound files. The interested user can view and hear a recording with a simple click in the text balloon that is associated to a placemark. At the same time transcriptions, translations, and other linguistic analysis layers based on the recording are displayed. Similarly, some multimedia lexica are made available through LEXUS [6] (Kemps-Snijders, Nederhof, & Wittenburg, 2006). After opening such a place-related link, a lexicon of the local language is opened.

Incorporating other web applications like VICOS[7], which aims at representing the knowledge domain of culturally relevant concepts, are interesting options for the future as well. Any other institute is welcome to add their own web applications pointing to such complex resource bundles.

Figure 3: The user gets direct access to multimedia-based annotations and lexica after clicking on a landmark. Data for the Yélî Dnye language (Rossel Island), provided by Levinson (2006).

[5] http://www.lat-mpi.eu/tools/annex
[6] http://www.lat-mpi.eu/tools/lexus
[7] http://www.lat-mpi.eu/tools/vicos

3. Representing Paths

A third application that was tested is related to language microvariation and language contact studies.

Recently, Levinson (2006) pointed out that certain geographical factors are much more relevant to miss language contact than others. Swamps, for example, are generally more difficult to cross than mountains and thus more frequently indicate a linguistic border. Therefore, it is common for researchers to study these linguistic microvariations: correlating cognate sets with geographic information, and so on.

From this perspective we would like to mention Burenhult (2008), who has used these GIS tools extensively while doing research on the Jahai language. He gathered GPS coordinates and has mapped his field sites to a GIS overlay file, i.e. a path can be shown to indicate at which time he was at which location. Each overlay file that is linked with nodes along this path contains references to the corresponding recordings, annotations, and field notes. In doing so, he can easily correlate his linguistic findings with geographic parameters. An example of this specific use can be found in figure 4.

In addition, today's GIS applications make it possible to add an arbitrary structure to the place mark collection in such a way, that anyone can easily turn on or off the visualization of certain phenomena. This results in a flexible representation enabling the researcher to focus on a very specific set of gathered data, e.g. showing only those locations where a specific linguistic feature occurs.

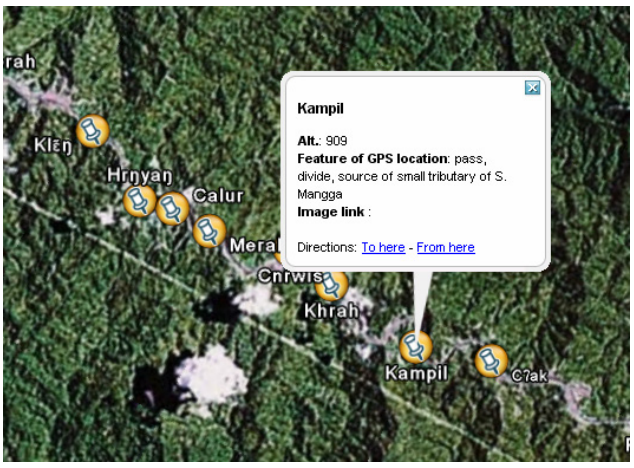

Figure 4: An example of how a path can be represented, including the reference to geographical features. Data for the Semang language (Malay Peninsula), provided by Burenhult (2008).

4. Anthropological Sites

In a fourth application material from anthropologists about mythical places, historically relevant events, and sociologically relevant material gathered over many years were extracted automatically from spreadsheets and transformed onto suitable overlays. Here, area indications or landmarks, which mark e.g. the movements of certain speaker groups or indicate where speakers talked about mythical stories were used and transformed to indicate a geographic area. Two illustrations of this are given in figures 5 and 6.

The unique possibility of zooming in and out offers completely new opportunities to the anthropologists in particular, when this information is superposed with material from other disciplines such as archeology (Beck, 2006), sociology (Matthews, 2003), etc.

Figure 5: The multidisciplinary approach: Linguistic glosses of the location name in combination with information on local living circumstances. Data for the Taa language (Namibia/Botswana), provided by Boden (2005).

5. Future research

A recent development in GIS systems is the addition of the time dimension to maps. Using an interactive time axis users can follow certain evolutions taking place on the depicted area. This could form an attractive way to present diachronic language changes. Another approach one could think of is adding geographical coordinates to each entry in a dialect lexicon. Afterwards, all of the variants can be easily plotted on a map, as illustrated by De Vriend et al. (2006) for Dutch dialect dictionary data.

6. Conclusion

Four possibilities to use geographic maps for linguistic research have been tested and we are sure that others will follow.

We have shown the strengths, but also the limitations of the aforementioned methods. They are particularly suited for research that correlates linguistic phenomena with geographic parameters and for the curious user who likes to virtually explore the surface of the earth. Linguists and anthropologists are yet just starting to experiment with these techniques. Other applications are imaginable, such as studying correlations between geographical and

linguistic features. It might also be a good opportunity to start building server-agnostic virtual collections, since the geographic access paradigm lends itself to surpass the traditional boundaries of individual language archives.

Most of the material mentioned above is already available on the web, in particular the Language-Sites overlay file can be downloaded. Some of the techniques mentioned are offered together with open material from the Yélî Dnye language spoken on Russell Island located at the South-East of New Guinea.

Figure 6: Toponyms and their etymology. This example features additional links directed to the local language community (*listen to the story*). Data for the Beaver language (Canada), provided by Miller (2003).

7. References

Beck, A. (2006). Google Earth and World Wind: remote sensing for the masses? *Antiquity, 80*(308).

Berck, P., Russel, A., Kemps-Snijders, M., & Wittenburg, P. (2005). *Advanced Web-based Language Archive Exploitation and Enrichment.* Paper presented at the 2nd Language & Technology Conference, Poznan, Poland.

Bird, S., & Simons, G. (2001). The OLAC metadata set and controlled vocabularies. *Proceedings of the ACL 2001 Workshop on Human Language Technology and Knowledge Management-Volume 15*, 7-18.

Boden, G. (2005). Western versus Eastern !Xõo – Difference, Politics, and Documentation. *Language Archives Newsletter, 1*(7), 2-7.

Broeder, D., Offenga, F., Willems, D., & Wittenburg, P. (2001). The IMDI Metadata Set, Its Tools and Accessible Linguistic Databases. *Proceedings of the IRCS Workshop on Linguistic Databases, Philadelphia*, 11-13.

Burenhult, N. (2008). Streams of words: Hydrological lexicon in Jahai. *Language Sciences, 30*(2-3), 182-199.

de Vriend, F., Boves, L., van den Heuvel, H., van Hout, R., & Swanenberg, J. (2006). *A Unified Structure for Dutch Dialect Dictionary Data.* Paper presented at the Proceedings of The fifth international conference on Language Resources and Evaluation, Genoa, Italy.

Goodchild, M. (2006). Geographic Information Systems. *Approaches to Human Geography*.

Kemps-Snijders, M., Nederhof, M., & Wittenburg, P. (2006). LEXUS, a web-based tool for manipulating lexical resources. *LREC 2006: fifth international conference on language resources and evaluation*, 1862–1865.

Levinson, S. (2006). Enrolling other sciences in language documentation: describing an isolate language in Papua New Guinea. DGFS Conference, Bielefeld.

Matthews, S. (2003). GIS and Spatial Demography. *GIS Resource Document*, 03-63.

Miller, J. (2003). *An Acoustic Analysis of Tone in Doig River and Blueberry River Beaver.* University of Washington.

Senft, G. (1986). *Kilivila: The Language of the Trobriand Islanders*: Walter de Gruyter.

Wittenburg, P., Mosel, U., & Dwyer, A. (2002). Methods of Language Documentation in the DOBES project. *Proceedings of LREC*, 34-42.