Visualization of Language Relations and Families: MultiTree

Damir Cavar, Malgorzata Cavar
Eastern Michigan University, The LINGUIST List
2000 Huron River Dr., Ypsilanti, MI 48197, USA
{damir,gosia}@linguistlist.org

Abstract

MultiTree is an NFS-funded project collecting scholarly hypotheses about language relationships, and visualizing them on a web site in the form of trees or graphs. Two open online interfaces allow scholars, students, and the general public an easy access to search for language information or comparisons of competing hypotheses. One objective of the project was to facilitate research in historical linguistics. MultiTree has evolved to a much more powerful tool, it is not just a simple repository of scholarly information. In this paper we present the MultiTree interfaces and the impact of the project beyond the field of historical linguistics, including, among others, the use of standardized ISO language codes, and creating an interconnected database of language and dialect names, codes, publications, and authors. Further, we offer the dissemination of linguistic findings world-wide to both scholars and the general public, thus boosting the collaboration and accelerating the scientific exchange. We discuss also the ways MultiTree will develop beyond the time of the duration of the funding.

Keywords: historical linguistics, language tree, visualization

1. About the project in general

MultiTree (Cavar et al., 2014) is an NFS-funded project collecting hypotheses about the genetic relationships between languages, dialects and language subgroups. The hypotheses are extracted from historical linguistics and typological literature. The hypothesis trees are stored in a database and linearized in form of a specific interoperable XML format. Some hypotheses about language relations are so complex that visualization as a tree is not possible, thus a graph representation with crossing relations is necessary. In two different publicly available web interfaces the language relation trees are displayed graphically in form of trees or graphs, using hyperbolic and general trees that can for example be interactively navigated, commented, searched, and compared.

Two online interfaces have been created over the last years. One online front-end is using a Java-based hyperbolic tree visualizer, which depends on a Java-plugin in the user's web-browser. The second interface is based on generated HTML and integrated JavaScript-components only. In the JavaScript-based interface general tree or graph visualization is realized with the D3 library (`d3js.org`) (Murray, 2013).

MultiTree visualizes a large number of linguistic theories, and offers a unique data base of language and dialect names, codes and bibliographic references. Currently more than 1,400 trees have been implemented, displaying the theories and models from more than 250 authors and an even larger number of publications. The MultiTree database is continuously extended and enriched, with more references and hypothesis trees added each month.

The MultiTree project has originally been funded in the years 2005–2009 and continues since summer 2012 till summer 2014.The database is on-line accessible in two versions. It offers free access to the language codes, names and references to the general public. While the original objective of MultiTree was to facilitate research in historical linguistics, it has developed into something much more powerful than a simple repository of scholarly information. In this presentation we would like to discuss the technical solutions we adopted while developing the new web interface, and the merits the project brings about beyond the research in historical linguistics.

2. Scholarly use

The original goal of the project was to collect the information about languages, language families, names and codes in a central database by extracting this information from numerous academic resources (e.g. books, research papers and dissertations). The resulting hypotheses about language relations are presented in the form of trees, which allows for an easy comparison and evaluation of different, complex, sometimes contradictory, hypotheses. The web page became over years a standard reference for students of linguistics but also for researchers. In its current new form it utilizes various visualization strategies that facilitate academic use of the collected material.

3. Codes

The same language may be in particular cases referred to using different names and/or different spelling conventions. In our database, languages, language groups and families are identified using their name AND a code, which allows for the unique and easy identification of a particular language regardless of the conventions employed within a scholar's publication. The project uses - whenever possible - standardized ISO 639-3 codes. This standard, however, is far from complete regarding languages, and it does not cover language groupings, or dialects. For items lacking an ISO code, the project team creates "local use" codes, which are used by other projects at the LINGUIST List, notably LL-Map, and are increasingly used by users external to the institution, e.g. Wikipedia.

The temporary language code is a unique three-character code randomly generated by the database, which, unlike

the ISO codes, can use both letters and numbers. For example, Pana (Pa-na, Bana) – Spoken in Hunan, China – a Hmong-Mien language with no ISO-code has been recently assigned a temporary code [3vs]. This particular language was listed in a classification by (Taguchi, 2013) and (Ratliff, 2010).

MultiTree dialect codes consist of three-letter ISO-code of the parent language, followed by a hyphen and a three-letter abbreviation of the dialect name. For example, Irish Gaelic [gle] has in MulitTree the following associated dialects: Munster dialect [gle-mun], Donegal dialect [gle-don], Connacht dialect [gle-con].

Local-use subgroup codes consist of four-letters. They include codes for established, well-documented, and relatively agreed-on language groupings: e.g. Austronesian [anes], Indo-European [ieur], Niger-Congo [ncon]. However, not all subgroups are well documented or agreed upon. When there are differences in the internal structure of subgroups, MultiTree will assign different subgroup codes. An example of this kind is the Northwest Formosan language subgroup [nwfr] (Li, 2006) vs. Northwest Formosan [bvqr] (Ethnologue, 2005) (spoken in Taiwan). Northwest Formosan in Ethnologue only has one child language, while Li's classification is much more complex, with eight languages and various sub-groupings with Northwest Formosan. Because of the radical differences between the two classifications, a new code must be created to avoid ambiguity.

At the moment, the database contains more than 20000 codes, including both ISO-standard and MultiTree-internal codes. Furthermore, the team members collaborate with the ISO-authorities, compiling documentation so that standard ISO codes can be assigned; for example, documentation for 352 languages, most of them Australian aboriginal languages, has been submitted between 2010-2013. MultiTree has been a contributor of the new code proposals to the ISO 639 family of standards.

When preparing submissions to ISO, MultiTree collaborates with the scholars whose focus of research was the language family or the subgroup that a case language belongs to. For instance, a change request for the new code element for a previously unrecognized Uralic language, Yurats, was prepared in cooperation with scholars Juha Janhunen and Tapani Salminen. Another example might be the change request for the existing ISO code [xgm]. In 2013 the documentation for the amendments to the codes [bjy] and [xgm]has been collected, to list Darumbal as the primary name connected with the code [xgm] (previously called Guwinmal), and change the scope of [bjy] code such that it does not cover Darumbal (together with Claire Bowern).

Existing	**Requested**
bjy Bayali, Darumbal	*bjy* Bayali
xgm Guwinmal	*xgm* Darumbal, dialect Guwinmal

Table 1: ISO change request for codes *bjy* and *xgm*

4. Language and dialect names

The use of multiple names with reference to the same language, dialect or group often makes it difficult for the audience to identify the topic of the scholarly endeavor, limiting its potential impact. By utilizing the unique codes, one can retrieve over our database a list of names for the particular language or dialect, and use it as a reference to disambiguate linguistic hypotheses.

4.1. References

By referring to the name of a language, or a dialect code, users gain access to a rich list of relevant references for a particular linguistic area. At the current moment, our publications database contains over 2000 entries and more than 200 authors.

4.2. International impact

The project web site is extensively used in the areas of the globe where the access to the primary sources is impossible or difficult, in Eastern Europe, and Africa, becoming an important and effective dissemination channel for scholarly efforts.

Apart from listing MultiTree as a reference in academic publications, we receive direct feedback from researchers. In just last year MultiTree team members have corresponded with MultiTree users, both scholars and language enthusiasts, in Cameroon, Liberia, Iran, South Africa, Indonesia, New Zealand, UK, Spain, the Netherlands, and Poland.

For example, MultiTree is often cited on the Russian Wikipedia on the pages for the languages that are lesser known, or endangered. The link to the MultiTree description of the code is given in the reference section of the Wikipedia page in Russian, while the English article of the same language often does not provide the same linkage. Examples of this kind include, among others, the Mod language of the Chadic language family, Tera languauge, Gorontalo language, Arawa language, etc. Similarly, Russian portal Garnish.ru, a collection of key knowledge, original ideas, useful services, features MultiTree on its page among general resources on linguistics and philology as one of the foreign multi-lingual and multi-ethnic portals. Another Russian portal for dictionaries and encyclopedias, Academic.ru, has several articles on the various languages which refer to MultiTree as one of the resources, for example, the article on Old Polish, or an article on the languages of (former) "Chechoslovakia." The unofficial website of the Department of Humanities of the St. Petersburg branch of the Higher School of Economics provides the reference to MultiTree as one of the key web resources for the Introduction to Anthropology courses.

5. Collaboration

A general moderated commenting facility has been integrated in the new MT-interface to enable collaborative research and commenting by experts. Commenting and discussion of trees and language relations is possible, as well as comments on codes and references.

For the moment the beta-version of the new incarnation of the web page uses Facebook and Disqus plug-ins for com-

ments about particular hypotheses, nodes on the trees or the trees themselves. In preparation is an interface for scholars who - after a password protected log-in will be able to add their own hypotheses or modify existing hypotheses by adding new subgroup nodes, new languages, new language names, codes, or simply additional references. This addresses one of the issues of the sustainability of the database after the funding runs out.

6. Technical issues

The first version of MultiTree uses Java based hyperbolic tree viewers, which allows to zoom-in on particular sub-sections of often very large trees containing thousand and more nodes. In this interface, available at the URL `http://multitree.org`, a search for language information can be specified by using language names or codes, or bibliographic references. In figure 1 we show the search result for the language code `hrv`. The Java-based hyperbolic-tree version of the interface is available at `http://multitree.org`. Figure 2 shows the selected references, (Brozović and Ilić, 1988), (Lončarić, 1996) that were taken as sources for the language tree, and the resulting tree for the language code `hrv-cha`, i.e. the Čakavian dialect of Croatian.

We created a second interface using the Python-based Django framework in the backend, and purely JavaScript-based frontend technologies for the general user interface. The user interface is using device adaptive technologies that are enabling common computing interfaces, as well as mobile devices and tablets. The tree visualization is achieved using only JavaScript libraries. Currently the visualization in the second interface is realized using the D3-library. All trees are stored in a database backend and made available as linearized XML-data-structures for sharing and dissemination.

Figure 3 shows a search for Pomoan in the database. The new interface displays references and the language tree (Campbell et al., 2007) in a similar way as the first Java-based interface. It allows, however, for scaling of the tree size using finger gestures on tablets or mouse-pads or various rotations of the tree.

7. Non-scholarly use

The web site is of interest not only for scholars. Site contents are presented in an accessible manner for all users. By clicking through trees, users can access information regarding unique languages, including extinction status or speaker numbers. Interestingly, speakers of various languages and non-professional language enthusiasts provide substantial feedback, which might be of use to the scholars.

8. Future: Crowd-sourcing

The final step of the MultiTree project is to develop the above mentioned scholars portal - an interface that enables scholars to enter into the database the hypotheses from their field of expertise without assistance of the MultiTree team members.The scholars portal hands over the main task of maintaining the contents of the database to the linguistic community world-wide, and reduces the institutional maintenance effort to the technical issues.

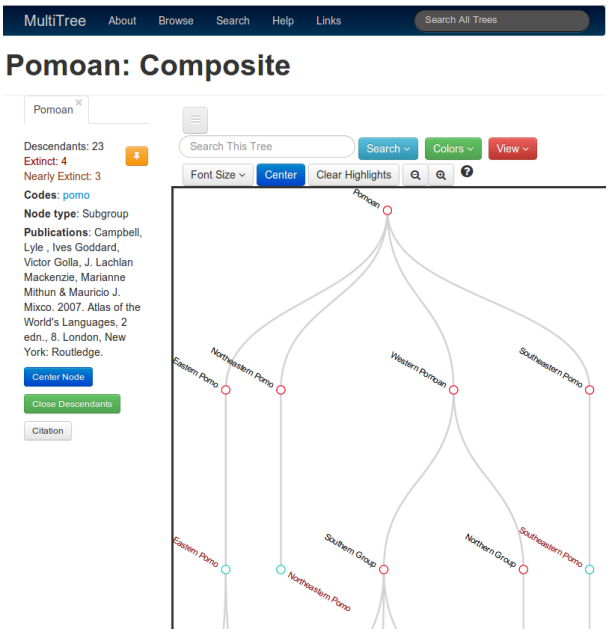

Figure 3: JavaScript-based tree view of the new MultiTree

9. Acknowledgements

The MultiTree project was funded by a National Science Foundation grant in years 2005-2009, and is continued since June 2012 (NSF grant no 1227106). We would lie to acknowledge the contribution of former PIs of the project, Anthony Aristar-Dry, Martha Ratliff, and Helen Aristar-Dry, students, interns, and The LINGUIST List staff who collaborated on the project, particularly, Lwin Moe, Eric Benzschawel, Myles Gurule, Uliana Kazagasheva, and Sara Couture.

10. References

Brozović, D. and Ilić, P. (1988). *Jezik, Srpskohrvatski/Hrvatskosrpski, Hrvatski ili Srpski. Izvadak iz II izdanja Enciklopedije Jugoslavije*. Jugoslavenski Leksikografski Zavod.

Campbell, L., Goddard, I., Golla, V., Mackenzie, J. L., Mithun, M., and Mixco, M. J., editors. (2007). *Atlas of the World's Languages*. Routledge, London, New York, 2nd edition.

Cavar, M., Cavar, D., Couture, S., Benzschawel, E., and Kazagasheva, U. (2014). Online visualization of research in historical linguistics. Poster presented at the LSA Annual Conference 2014, January.

Lončarić, M. (1996). *Kajkavsko narječje*. Školska knjiga, Zagreb.

Murray, S. (2013). *Interactive Data Visualization for the Web: An Introduction to Designing with D3*. O'Reilly Media, March.

Ratliff, M. (2010). *Hmong-Mien language history*. Pacific Linguistics, Canberra.

Taguchi, J. (2013). On the phylogeny of hmongic languages. Presented at 23rd Annual Meeting of the Southeast Asian Linguistics Society, May.

Figure 1: Search results for language code `hrv`

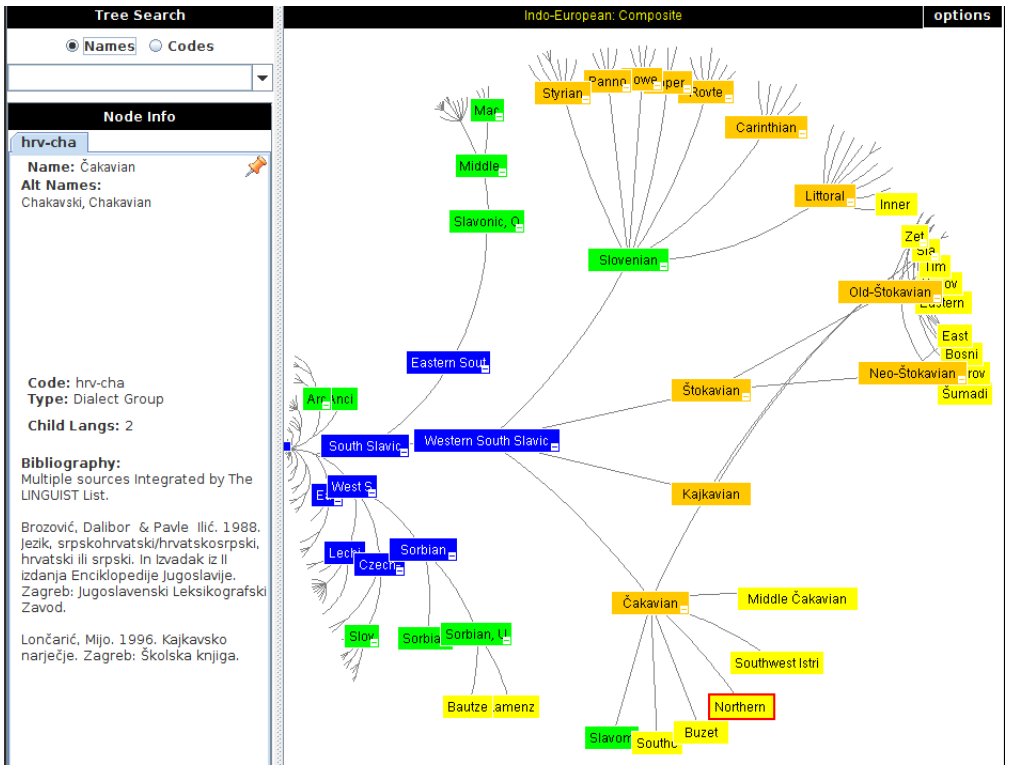

Figure 2: Java-based hyperbolic tree view of MultiTree