Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/978-1-61499-442-8-192

Compilation of Language Resources and On-Line Dissemination of Knowledge about Endangered Languages and Linguistic Heritage

Katarzyna KLESSA[a,1] and Nicole NAU[b]
[a]*Department of Phonetics,* [b]*Department of Baltic Studies*
[a, b]*The Institute of Linguistics, Adam Mickiewicz University in Poznań, Poland*

Abstract. This paper presents selected results of two projects concerning compilation of language resources and dissemination of knowledge about endangered languages and language documentation primarily among non-academic target groups. Both projects include data from one of the Baltic languages, Latgalian. The aim of the paper is to discuss the main features of the resources as well as to show the interest of these projects from a Baltic perspective and their usefulness for further research and development.

Keywords. Compilation of language and speech resources, knowledge dissemination, endangered languages, Latgalian

Introduction

For the last two decades, the documentation of endangered languages has been one of the key topics in international linguistics, and documentary linguistics has become a meeting ground for researchers from diverse domains such as field linguistics, descriptive linguistics, and language technology. Documentation projects carried out according to modern standards have resulted in considerable collections of text, audio and video data of hitherto little known languages. A challenge for the present and the future is to make the best use of these data for research as well as for the benefit of the speech communities. A further group of potential users is the general public, who is still little aware of the current linguistic diversity and the imminent danger of its disappearance. On this background, two projects carried out at Adam Mickiewicz University in Poznań, Poland, have developed platforms for the dissemination of knowledge about endangered languages and their documentation.

[1] Corresponding Author: Katarzyna KLESSA, Department of Phonetics, The Institute of Linguistics, Adam Mickiewicz University in Poznań, Poland; E-mail: klessa@amu.edu.pl

1. Languages in Danger: Dissemination and Education

The interactive website at http://languagesindanger.eu/ delivers information about selected endangered languages from all over the world, and is designed for students and teachers of secondary schools as well as for young researchers with a view to increase awareness and interest in these topics [1]. The implemented features include: interactive components (interactive map, interactive quizzes), teaching materials (ready-to-use classroom materials, multimedia resources), a textbook for background reading (10 chapters enriched with visual and multimedia material). The components were created using source data available from existing repositories e.g. DoBes [2] as well as with the support of new data gathered by individual researchers and language documenters.

More than 20 languages are selected as featured languages of the website. Locations of these featured languages are displayed in the interactive map together with exercises and multimedia materials. They are also referred to in more detail in other sections of the website. Among others, these featured languages include Latgalian [3] (spoken in Latvia), Karaim (spoken in Lithuania and Poland), Wilamowicean [4] (spoken in Poland). There are also teaching materials about Finno-Ugrian languages, which could be especially interesting for Estonian schools.

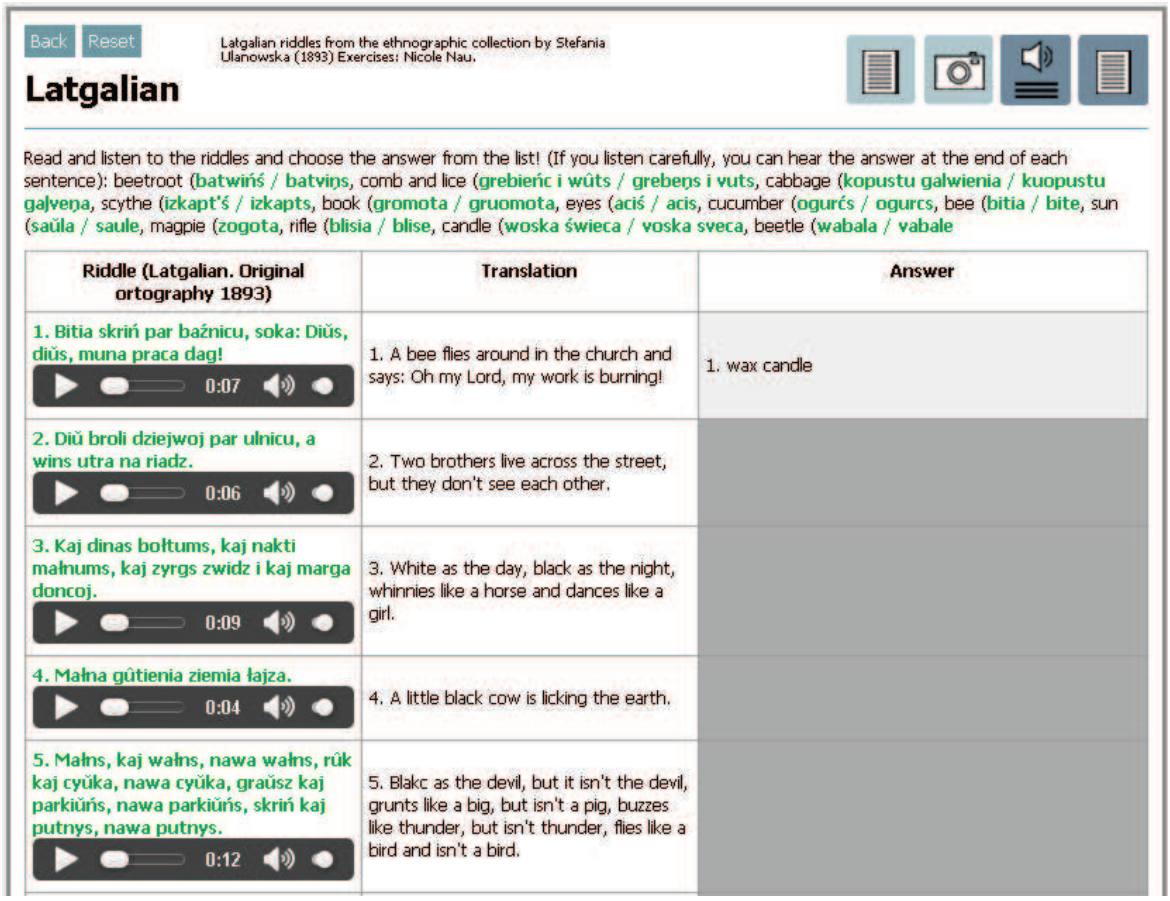

Figure 1. Latgalian exercise on the Interactive Map (the task involves listening and reading).

An important novelty of the package is the European perspective and the availability of all contents not only in English but also in Polish, Hungarian, German and Dutch (to be published in September 2014). The language versions of the website are creative adaptations rather than just direct translations [5]. The *Languages in Danger* website was developed using WordPress Open Source web software [6]

2. Linguistic Heritage: Compilation of Language Resources and Dissemination

The main goals of the project *Dziedzictwo językowe Rzeczypospolitej* (The linguistic heritage of the Polish-Lithuanian Commonwealth, referring to the languages spoken in the territory once named 'Rzeczypospolita' in Polish) were:

- to provide information about languages for a general audience (above 20 languages described so far, including Baltic (Latgalian; Lithuanian dialects spoken in Poland) and other languages also spoken in the Baltic States (e.g. Yiddish, Romani, Karaim, Russian dialects of Old Believers).
- to start a scientific documentation database providing resources for researchers (linguists) (4 languages featured so far: Latgalian, (Polish) Yiddish, Wilamowicean, Halcnovian).

Latgalian resources in the on-line database (access at: http://inne-jezyki.amu.edu.pl/) contain among others folklore texts collected by Stefania Ulanowska in the 1890s. The original edition contained the Latgalian texts written in the semi-standardized writing system of the time and a Polish translation. In the present database, a new transliteration based on today's orthography was added. Moreover, for all the original fairy tales (and selected songs and short tales) English translations have been provided. For some of the tales, audio recordings were made and also included in the database.

The Linguistic Heritage database can be flexibly extended by new languages and source data. Future works could include: a) adding resources from Lithuanian dialects spoken in Poland; b) providing material from contemporary Latgalian; c) further development of functionality and interface (configurable search and display, data export, embedded use of annotation tools, e.g. Annotation Pro [7]).

The Linguistic Heritage database was implemented with SQL Server technology. To enable simultaneous data edition by multiple users, an on-line data-management application was developed using ASP.NET. The multimedia files were annotated with: Elan [8] (video), and Annotation Pro [7] (audio).

3. Final Remarks

The architecture of both of the described websites enables further extensions by other language versions (adaptations) and materials, e.g. with the participation of a potential new partner(s) from a Baltic State. For the Linguistic Heritage Database tools have to be developed that meet the needs of linguists and ethnologists who want to use the data in their research.

It is important to realize that language documentation does not end when data has been collected and stored in digital archives. These archives are more than museums of dying languages, they provide powerful resources for further research.

Acknowledgements

This work was supported by **INNET** (*Innovative networking in infrastructure for endangered languages*), European FP7 grant agreement no: 284415, in the years 2011-2014, and by the Polish Ministry of Science and Higher Education within "The National Programme for the Development of Humanities" in the years 2012-2013.

References

[1] Wójtowicz, R. 2014. Language endangerment in European secondary schools: Challenges and perspectives. In*: Proceedings of the International Conference The Future of Education IV*, Florence. On-line at: http://conference.pixel-online.net/FOE/conferenceproceedings.php]

[2] DOBES (Documentation Of Endangered Languages). Available on-line at: http://dobes.mpi.nl/

[3] Nau, N. 2011. *A short grammar of Latgalian*. München: LINCOM Europa.

[4] Wicherkiewicz, T. 2003. *The Making of a Language. The case of the idiom of Wilamowice*, Southern Poland. Berlin-New York: Mouton de Gruyter.

[5] Jung, D., Klessa, K., Duray, Z., Oszkó, B., Sipos, M., Szeverényi, S., Várnai, Z., Trilsbeek P., Váradi, T. 2014. Languagesindanger.eu - including multimedia language resources to disseminate knowledge and create educational material on less-resourced languages. In: *Proceedings of the 9th Language Resources and Evaluation Conference (LREC),* Reykjavik.

[6] Klessa, K., Karpiński, M., Wagner, A. 2013. Annotation Pro - a new software tool for annotation of linguistic and paralinguistic features. In: *Proceedings of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop*, Aix en Provence (Software download site at: http://annotationpro.org/downloads/).

[7] *WordPress Open Source web software*. Available on-line at: https://wordpress.org/

[8] Sloetjes, H., Wittenburg, P. 2008. Annotation by category – ELAN and ISO DCR. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC),* Marrakech.