Technical and social Infrastructures for the Humanities: The Example of the Dagaare-English-Cantonese Dictionary

Bodomo, Adams
adams.bodomo@univie.ac.at
Universität Wien, Institut für Afrikawissenschaften; AT

Wandl-Vogt, Eveline
eveline.wandl-vogt@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Austrian Centre for Digital Humanities; AT

Mörth, Karlheinz
karlheinz.moerth@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Austrian Centre for Digital Humanities; AT

Introduction

This paper introduces into the transformation process of the Dagaare – Cantonese – English dictionary into an open, online research infrastructure in the framework of European research infrastructures and – in doing so – open those for Non-European researchers, research data as well as topics.

The trilingual dictionary is designed for use in lexicographical and linguistic field methods training. It serves as a database to illustrate many linguistic principles and phenomena in phonology, morphology, syntax and semantics. First and foremost it is intended as a reference source for Chinese and English speaking students.

Dagaare is a language spoken in Ghana and Burkina Faso by about two million people. It belongs to the Gur branch of the Niger-Congo family. In spite of the fact that Dagaare is genetically unrelated to Chinese, there are some interesting typological features under which the two languages can be compared. To illustrate, both Dagaare and Chinese are tone languages, unfortunately lacking audio files in the printed dictionary version. But while Chinese has a complex system of four to nine tonemes, Dagaare - like most West African languages - has a two-tone system. The first part of the dictionary includes information of the orthography and sound system of Dagaare followed by an explanation of the verbal and nominal morphology of this language. Part two is the proper dictionary which comprises more than thousand head words and a total of 3,000 to 4,000 words. Subsequent to the lexicon are represented sample field work projects that are intended to aid both the field trainer and trainee. They cover the areas of phonology, morphosyntax, lexical semantics ans sociolinguistics.

The valuable lexicographical data mentioned before were meant to be made sustainable available on the internet. To this end, they had to be transferred into an existing infrastructure, the research infrastructure for lexicography available at the Austrian Centre of Digital Humanities (ACDH) of the Austrian Academy of Sciences. The Academy has a long-standing tradition in eLexicography to which several departments contributed over more than hundreds of years. Most recently, the ACDH hosts a research group on eLexicography, the lexicography laboratory (1.1.2015-), to support, coordinate and methodically explore experimental scholarship in the fields of lexicography.

The emerging infrastructure is made up of several components: (1) an editor, (2) a formalised encoding framework, (3) a depositing back end and (4) a publishing system, all of which have been integrated into one system. An important keyword in this endeavour has been modularisation, the system not being one single piece of software but a number of complementary components that interlock neatly through clearly defined interfaces.

Workflow

The work of integrating the lexicographical data into the infrastructure was performed by the eLexciography working group of the ACDH. The workflow is a five step procedure:

(1) analysing and discussing the research focus and data structure

(2) converting the data from a simple table into a standards-based XML format (TEI P5)

(3) importing it into the database

(4) manual post processing

(5) and publication on the internet.

At the end of the process the data will be available in a persistent manner.

Editor

The Viennese Lexicographical Editor (VLE) is a fairly new piece of software that first came into existence as a by-product of an entirely different development activity: the creation of an interactive online learning system for university students. Thus, it was first used in a collaborative glossary editing project carried out as part of university language courses at the University of Vienna. As the tool proved to be flexible and adaptable enough, it was also used and further developed in a number of other projects collecting lexical data.

The interface is built around an XML editor that allows to process standard-based lexicographic and terminological data. Basically any XML-based formats such as LMF, TBX, RDF or TEI can be handled. The program provides a number of useful functions to automate editing procedures. It can check the structural integrity (well-formedness) of input on the fly. Technologically, it draws not only on the XML core specification but also on several cognate technologies. XSLT and XPath play an important role both for visualising and modifying existing datasets. Lexicographers can insert elements on the basis of predefined XML schemas. Most of the functions can be applied both to single and multiple lemmas. One of the most recent improvements is a versioning system and an improved working mode that allows lexicographers to work on the XML data without actually seeing the tags. Furthermore, the editor also has a configurable interface enabling lexicographers to access external corpora and to integrate example sentences from them into dictionary entries. The communication between the dictionary client and the server has been implemented as a RESTful web service.

The tool forms part of Austria’s contribution to the pan-European CLARIN Research Infrastructure Consortium and is freely available from the ACDH Website .

Formalised encoding framework

While the list of formats used in the lexicographic community is unfortunately very long, there exists a de-facto standard which has been used widely in many digital humanities projects, in numerous lexicographic projects and most of the ACDH’s lexicographic endeavours: the Guidelines of the Text Encoding Initiative (TEI). The application of digital (de-facto) standards in building digital language resources is of particular concern when we think about interoperability and re-usability of resources. The buzz-word of open life-cycles for research data will remain meaningless unless researchers succeed in achieving a certain degree of harmonisation in structuring their data and meta data. The ACDH has been working on specialised schemata based on the TEI (P5) dictionary module for quite some time. In all these efforts, they have also aimed at a high degree of interoperability with the ISO standard LMF (Lexical Markup Framework). In order to realise mechanisms for cross-dictionary access, they have also been working with semantic technologies such as RDF and SKOS.

The basic reduced TEI schemata have been documented in form of guidelines which give detailed accounts of how dictionaries in the ACDH collection were encoded. These guidelines document and discuss the schema and furnish a number of examples taken from actual dictionaries. The target group for this guide are both the lexicographers working on ACDH projects as well as others who might want to work along similar lines. These particular Guidelines were themselves produced making use of the TEI framework.

Depositing infrastructure

The dictionary editor is a web-based application that allows lexicographers to work in groups. The data is stored on a server of CLARIN Centre Vienna. Being part of an official infrastructure, long-term availability will be vouchsafed. In addition, the owners of the lexicographical data can draw copies of their dictionaries at any time of the compilation process.

Publishing framework

The publishing infrastructure builds on corpus_shell, a service-oriented architecture and a distributed and heterogeneous virtual landscape. The core functionality of this modular framework is to expose well-defined interfaces based on acknowledged standards. The principle idea behind the architecture is to decouple the modules serving data from the user-interface components. One of the nice features of the system is that you can build new interfaces via XSLT styles almost on the fly.

Status and outlook

The conversion and import of the data has already been undertaken. Dagaare – English – Cantonese Dictionary is already available online. However, the working group is still improving the web-interface, a stable URL will be assigned by the end of 2015.

The Dagaare – English – Cantonese Dictionary is about to be improved from a content as well as collaboration / social infrastructure point of view:

(1) Audio files are to be added to support – mainly – the representation of the tone languages Dagaare and Cantonese. Doing so, we enlarge the network of people participating into the project for both,

(a) free and open Wikimedia audio tools as well as

(b) high performance audio tools e.g. supported by Forschungszentrum Telekommunikation Wien (FTW http://www.ftw.at/) or Phonogrammarchiv at the AAS http://www.phonogrammarchiv.at/ ,

(c) speakers with Cantonese mother tongue. (Bodomo Adams himselves represents Dagaare mother tongue).

(2) The dictionary will be fully embedded into the lexicographical research infrastructure of ACDH as well as the research of the Institut für Afrikawissenschaften at the University of Vienna. This implies both,

(a) experimental development of the dictionary content applying methods of other disciplines e.g. Natural Language Processing and Semantic Technologies for interlinking with other dictionaries, semi-automatic translation into other languages starting with German, connecting with cultural content etc., e.g. interlinking with cultural resources like songs; (b) embedding it into a research framework for African Diaspora studies.

In doing so, the representatives of both institutes open towards collaboration of global communities of several disciplines that are until now not in touch.

Bibliographie

Bodomo, Adams (2004): *Dagaare – Cantonese – English Dictionary for Lexicographical Field Research Training* (= Afrikawissenschaftliche Lehrbücher 14). Köln: Köppe.

Bodomo, Adams / Mora, Manolete (2007): "Documenting Spoken and Sung Texts of the Dagaaba of West Afrika", in: *Empirical Musicology Review* 2, 3: 81-102.

Budin, Gerhard / Majewski, Stefan / Moerth, Karlheinz (2012): "Creating Lexical Resources in TEI P5", in: *Journal of the Text Encoding Initiative* 3 https://jtei.revues.org/522 [letzter Zugriff 08. Februar 2016].

Budin, Gerhard / Moerth, Karlheinz (2011): "Hooking up to the corpus: the Viennese Lexicographic Editor's corpus interface",in: Kosem, Iztok / Kosem, Karmen (eds.): *Electronic lexicography in the 21st century*. New applications for new users. Proceedings of eLex 2011 conference. Bled, Slovenia: Trojina, Institute for Applied Slovene Studies 52-59.

Budin, Gerhard / Moerth, Karlheinz / Durco, Matej (2013): "European Lexicography Infrastructure Components", in: Kosem, Iztok / Kallas, Jelena / Gantar, Polona / Krek, Simon / Langemets, Margit / Tuulik, Maria (eds.): *Electronic lexicography in the 21st century: thinking outside the paper*. Proceedings of the eLex 2013 conference, 17-19 October 2013. Tallin, Estonia: Trojina, Institute for Applied Slovene Studies / Eesti Keele Instituut 76-92.

Declerck, Thierry / Lendvai, Pirsoka / Moerth, Karlheinz (2013): "Collaborative Tools: From Wiktionary to LMF, for Synchronic and Diachronic Language Data", in: Francopoulo, Gil (ed.): *LMF*. Lexical Markup Framework. London / Hoboken: John Wiley & Sons 175-186.

Declerck, Thierry / Moerth, Karlheinz / Wandl-Vogt, Eveline (2014): "A SKOS-based Schema for TEI encoded Dictionaries at ICLTT", in: *LREC 2014, Ninth International Conference on Language Resources and Evaluation*. Reykjavik, Iceland: European Language Resources Association 414-417.