Cloud Computing for Linguists

Dorothee Beermann
Norwegian University of Science and Technology
dorothee.beermann@hf.ntnu.no

Pavel Mihaylov
Ontotext
pavel@ontotext.com

Abstract

The system presented is a web application designed to aid linguistic research with data collection and online publishing. It is a service mainly for linguists and language experts working with language description of less-documented and less-resourced languages. When the central concern is in-depth linguistic analysis, maintaining and administering software can be a burden. Cloud computing offers an alternative. At present mainly used for archiving, we extend linguistic web applications to allow creation, search and storage of interlinear annotated texts. By combining a conceptually appealing online glosser with an SQL database and a wiki, we make the online publication of linguistic data an easy task also for non-computationally oriented researchers.

1 General description of TypeCraft

TypeCraft (or TC in short) is a multilingual online database of linguistically-annotated natural language texts, embedded in a collaboration and information tool. It is an online service which allows users (projects as well as individuals) to create, store and retrieve structured data of the kind mainly used in natural language research. In a system featuring graded access the user may create his own domain, invite others, as well as share his data with the public. The kernel of TypeCraft is morphological word level annotation in a relational database setting, wrapped into a wiki which is used as a communication and information gathering and sharing tool. TypeCraft allows the import of raw text for storage and annotation and export of annotated data to MS Word, OpenOffice.org, LaTeX and XML. The online system is complemented by an offline client which is a Java application offering the same functionality as the online version. This allows a seamless exchange of data between the server and the user's own computer.

2 Online system internals

The online system is supported by a central server running the following modules: TypeCraft server proper, an SQL database, Apache, MediaWiki. The client side consists of the TypeCraft editor interface and a wiki environment (content produced by MediaWiki on the server). Users perceive the wiki and the editor interface as a single TypeCraft web application.

The TypeCraft server proper is a Java application running inside a Java application server. TypeCraft uses a PostgreSQL database for data storage. The data mapping between Java objects and database tables is managed by Hibernate, so the system is not bound to any specific SQL database. TypeCraft data can be divided into two distinct groups: *common data*, shared between all annotated tokens and users, such as the word and sentence level tag sets and an ISO 639-3 specification, and *individual data*, by which we mean specific texts, phrases, words and morphemes. Individual data references common data types. This for example means that all users of the system making use of the part of speech tag N share the reference to a single common tag N.

3 Digital linguistic data

It is well known that generation of linguistic annotation of any kind is a time consuming enterprise quite independent of the form the primary data has and the tools chosen for processing this data. Equally well known are problems connected to the generation and storage of linguistic data.

Coling 2010: Demonstration Volume, pages 49–52,
Beijing, August 2010

Standard word processing programs do not function well as linguistic tools and private computers are not a safe place to store linguistic resources (Bird and Simons, 2003). Although it is generally agreed that linguistic resources should be kept in a sustainable and portable format, it is less clear what that really means in practice. For the individual researcher it is not easy to decide which of the available tools serve his purpose best. To start with, it is often unclear which direction the research will take, which categories of data are needed and in which form the material should be organised and stored. We experience that it is too time consuming or requires expert knowledge to convert otherwise useful data into an acceptable 'resource format'. It is perhaps even more important that many tools turn out to be so complex that the goal of mastering them becomes an issue in its own right. Researchers working together with local communities on less-documented languages experience that linguistic software can be technically too demanding.

In fact, researchers in all non-computational fields of linguistics encounter problems similar to those just described for field-oriented research. Concerned with timely publication, for which linguistic data mainly takes the form of Interlinear Glosses (IG), the efficiency with which linguistic data can be created is an important issue. Several factors will affect which form linguistic data management will take, namely the standardisation of data beyond the field of NLP, non-expert user IT solutions allowing the efficient creation of linguistic data, and finally, improved availability of linguistic data for human consumption in research and publication.

4 Linguistic services and public linguistic data

Within linguistics the idea of cloud computing is relatively new: the basic concept is that users of digital technology no longer need to maintain the software they use, instead the maintenance of the technological infrastructure is left to services online. Already a success in commercial applications, IT services have also become a reality in research. Within linguistics and specifically language documentation, cloud computing facilities are at present mainly restricted to online archives. Yet, online services can be extended to provide tools for databasing and annotation of data. Scientific data exchange is an issue in biochemistry (Leser, 2009), but as far as we know it has not been an issue in linguistics. The question is not so much why we should share data but rather *how* and *what*. The linguistic tool that we would like to demonstrate gives a concrete answer to these questions. Table 1 presents a short overview of the main functionalities of the TypeCraft web application.

5 Creation, storage, migration and representation of IGs in TypeCraft

The TypeCraft web application can be used online at `http://www.typecraft.org/`. The TC wiki serves as the central hub of the application. The TC database is accessed through *My Texts* which displays the user's repository of IG collections, called *Texts*. *My Texts* is illustrated in Figure 1. Graded access is one of the design properties of TypeCraft. *My Texts* has two sections consisting of private data (data readable only be the user), and shared data. Shared data are Texts owned by groups of TC users. After being assigned to a group, the user can decide which data to share with which of his groups. Data can also be made public so that anyone on the net can read and export (but not edit) it.

TypeCraft is like the well known Linguist's Toolbox (International, 2010) an interlinear glosser. However, different from Toolbox, TypeCraft is a relational database and therefore by nature has many advantages over file-based systems like Toolbox; this not only concerns data integrity but also data migration. In addition, databases in general offer greater flexibility for search and retrieval. The other major difference between Toolbox and TypeCraft is that TypeCraft is an online service which frees the users from all the problems arising from maintaining an application on their own computer. Online databases like TypeCraft are multiuser systems, i.e. many people can access the same data at the same time independently of where they are located. Users administer their own data, either in a private domain or publicly, and they can make use of other users'

Table 1: Overview over TypeCraft Functionalities

Annotation	Collaboration	Data Migration
sentence tokenisation	graded access	manual text import
interactive table cells	tool internal user communication	export of annotated phrases to MS Word, OpenOffice.org and LaTeX
Lazy Annotation Mode	user pages for background information	XML semi-automatic export to the TC wiki
extensive search functionality	sharing of data sets between user groups	automatic update of data exported to the TC wiki

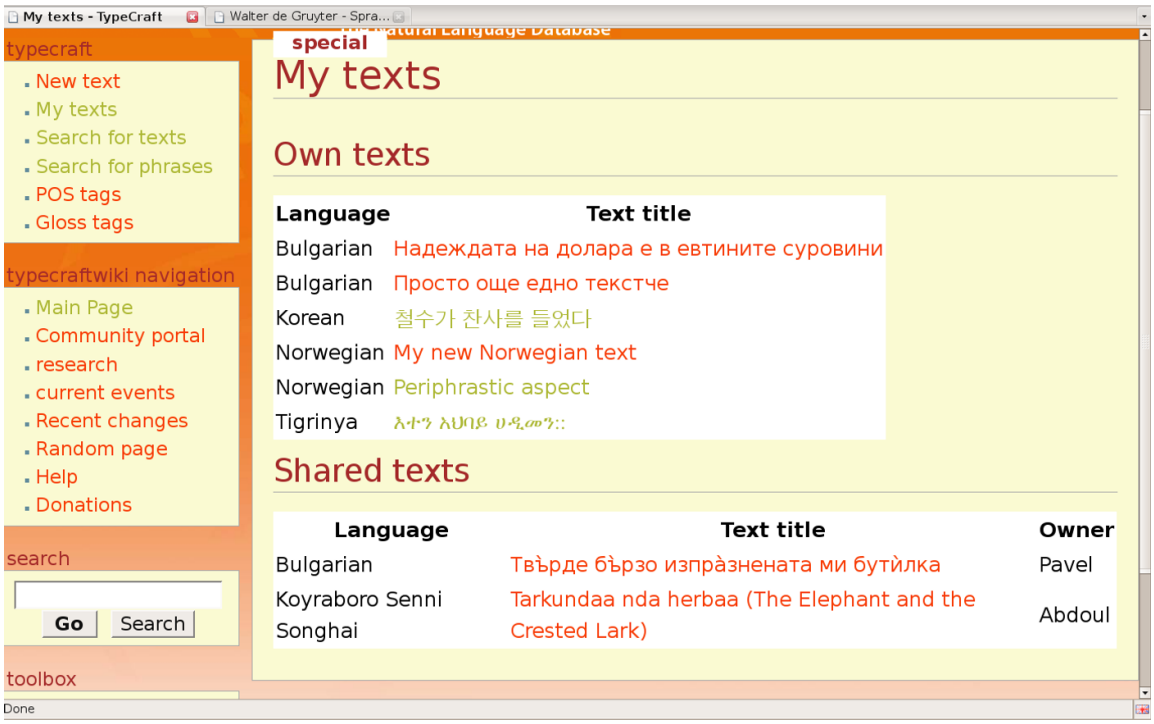

Figure 1: My texts in TypeCraft

data. Sharing information and data is an issue of mutual interest. Using standard wiki functionality, users discuss annotation issues. A TC internal email function allows users to communicate directly within the application. User pages function to personalise information and to create a TC user community. Social networking within a scientific tool plays a crucial role for the improvement of data validity. Information given by annotators, such as native language and professional background, increase the trust in TC data.

The TC wiki features interactive Google maps (a MediaWiki extension) which can be used to locate a language geographically. Isoglosses can be shown on the map too.

It is not always possible to work online. The TC online database is complemented by the TC offline client which can be downloaded from the project website for free. As a Java application it runs on multiple platforms, and allows the user to work offline in an environment familiar to him from the web application. The offline client offers the same functionality as the online service. The user can import data either locally or from the central TC database.

6 Glossing with TypeCraft

TypeCraft supports word-to-word glossing on eight tiers. After having imported a text and run it through a simple sentence splitter, the user can

click on a phrase and enter annotation mode. The system prompts the user for *lazy annotation* (in Toolbox called *sentence parsing*) which will automatically insert the annotation of already known words into the annotation table.
The user is restricted to a set of predefined tags which can be accessed from the TC wiki navigation bar where they are automatically updated when the database changes. TypeCraft is a multilingual database hosting languages from distinct language families and grammar traditions. It is therefore crucial to have standards that are extendible.

The TypeCraft tag set is mapped to the General Ontology for Linguistic Description (GOLD). GOLD (Farrar and Langendoen, 2003) has been created to facilitate a more standardised use of basic grammatical features. As an OWL ontology GOLD allows a representation of grammatical features in terms of categories and their relations. By mapping TC tags to GOLD, the user can make use of the information in the GOLD system which allows him to relate tags to more general grammatical concepts. The TypeCraft–GOLD mapping allows the user direct access to standards and necessary background information to associate glosses with the grammatical categories they are meant to express. GOLD in many cases provides definitions of concepts and important bibliographic resources related to the use of the term.

Annotated TC tokens can be exported to Microsoft Word, OpenOffice.org Writer and LaTeX. Example (1) is exported to LaTeX from TypeCraft. It illustrates locative relativisation in Runyakitara, a Bantu language spoken in Uganda:

(1)

Omu nju ei abagyenyi baataahiremu ekasya						
òmù	njù	èì	àbàgyènyì	bàátàhìrèmù		èkásyà
Omu	n ju	ei	a ba gyenyi	ba a	taah ire mu	e ka sya
in	CL9 *house*	*which*	IV CL2 *visitor*	CL2 PRS.PERF *enter*	PERF LOC	CL9 PST *burn*
PREP	N	REL	N	V		V
'The house in which visitors entered burned'						

Next to export to the main text processing systems, TypeCraft supports XML export which allows the exchange of data with other applications.

7 Conclusion

Interlinear Glosses are the most common form of linguistic data annotated by humans. In this paper we have presented an online linguistic service which allows the creation, storage and retrieval of IGs, thus granting them the status of an independent language resource. Reusability of data has become an issue also in the non-computational fields of linguistics. Although not sufficiently rewarded at the moment, already now the creation and sharing of linguistic data online is an efficient way for the creation and propagation of annotated texts in form of Interlinear Glosses. Since the TypeCraft web application provides off-the-shelf data for linguistic publications already formatted for all main text processing systems, data creation and retrieval with TypeCraft is time efficient. This makes linguistic work more data oriented and enables reasonable scientific turnover rate.

References

Bird, Steven and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Languages*, 73(3):557–582.

Farrar, Scott and D. Terence Langendoen. 2003. A linguistic ontology for the semantic web. *GLOT International*, 7(3):97–100.

International, SIL. 2010. http://www.sil.org/, January.

Leser, Ulf. 2009. Social issues in scientific data exchange. Manuscript Humboldt Universität, Berlin.