TypeCraft: Collaborative Databasing and Resource Sharing for Linguists

Dorothee Beermann[1] and Pavel Mihaylov[2]

[1] Norwegian University of Science and Technology, Trondheim, Norway
dorothee.beermann@hf.ntnu.no
[2] Ontotext, Sofia, Bulgaria
pavel@ontotext.com

Abstract. We present a linguistic application that uses web technologies to promote the reuse of research data in the form of Interlinear Glossed Text (IGT), which is a well-established data format within philology and the structural and generative fields of linguistics. Here we present the modules and procedures of the online database TypeCraft.[3] IGT is a sought after commodity in NLP and an integral part of scholarly linguistic work. It not rarely represents the only structured data available for less-resourced or endangered languages. While archiving of structured data from endangered languages is already well on its way [2], the free creation and exchange of linguistic data in the form of linked IGTs still needs to gain in popularity.

1 Introduction

With more linguists interested in the description of endangered and less-described languages, we see a rising interest in the electronic creation and management of linguistic data. Distinct from computational approaches to language annotation, where tagging of language data serves to facilitate automatic processing, annotation by linguists is meant for human consumption. TypeCraft (TC), the tool we are going to present here, is specialised on the creation, management and exchange of Interlinear Glossed Text (IGT), exemplified by an example from Runyankore Rukiga (ISO 639-3,nyn), a Lacustrine language of the Great Lakes area in Uganda:

In: C. Unger, P. Cimiano, V. Lopez, E. Motta, P. Buitelaar, R. Cyganiak (eds.): Proceedings of Interacting with Linked Data (ILD 2012), Workshop co-located with the 9th Extended Semantic Web Conference, Heraklion, Greece, 28-05-2012, published at http://ceur-ws.org

[3] www.typecraft.org

Omu nju hakataahamu abagyenyi
òmù njù hàkàtààhàmù àbàgyéngyì
Omu n ju ha ka taah a mu a ba gyenyi
in CL9 *house* CL16 PST *enter* IND LOC IV CL2 *visitor*
PREP N V N
'In the house entered visitors'

At present the main hindrance for IGTs to be a prime linguistic resource is the still prevalent lack of glossing standards. In-depth manual linguistic annotation is time consuming and often cyclic since it is an integral part of a scientific discovery process. A tool facilitating this process needs to allow for partial annotation It further needs to take into consideration that linguistic annotation is often distributive work done in close collaboration between native speakers and linguists, so that a tool that can be used independently by all partners has to cater to the needs of users with different levels of linguistic expertise and computer literary.

The discussion is organised as follows: In Section 2, we present the TypeCraft system from a user perspective. We discuss the creation and management of IGT and outline how collaborative online databasing can contribute to making linguistic data re-usable and mobile. In Section 3, we give a system description and in Section 4, we will summarise, and outline future work.

2 TypeCraft

TypeCraft (TC) [1] consists of a relational database combined with a tabular text editor for the computer-aided manual annotation of text. The tool's outer wrapper is a customised mediawiki which serves as a general entrance port and collaboration tool. The list below gives an overview over TC's main functionalities:

Annotation
Manual import of continuous text or sentence collections
Tabular interface for word level glossing with automatic sentence break-up and direct access to pre-defined gloss-sets
Easy access to gloss definitions and linking to an online ontology of grammatical concepts
Data management
Wiki facilitated management of primary data and metadata
Easy navigation between primary and annotated data
Easy linking of data to the TC-wiki for an integrated multi-media representation of data
Collaboration
Graded access and individual work spaces: the user decides which data is shared, TC texts and phrases have their own URL and thus can be acquired and exchanged freely online
User groups can share data
Co-editing of TC-wiki pages for early dissemination
Data export
Export of annotated sentences (individuals or sets) to Microsoft Word, Open Office and LaTEX for paper publications. Print-friendly versions of the TypeCraft web pages including exported database material
Export of XML for automatic data processing

2.1 General Information

While other manual annotation tools designed for normal linguists are desktop systems,[4] TC is a multi-user online system. Users administer their own data in

[4] The other tools are Toolbox and FLEx, both developed by SIL.

their private space at TC, but they can also make use of other users' shared data. TC is loaded by directing a browser[5] to www.typecraft.org. We use standard wiki functionality to direct the user to the database via *New text*, *My texts*, and *Text- or Phrase search*. *My texts* displays the user's repository of annotated material, called 'Texts'. *My texts*, the user's private space, is divided into two sections: *Own texts* and *Shared texts*. This reflects the graded access design of TC. Texts can be shared between groups. Data-sets can be easily exchanged between users and co-edited in user-groups. Interlinear Glosses can be loaded to the TC-wiki where they are displayed as part of a TC-wiki page. Under the printing of wiki pages TC data export automatically. One additional feature is that wiki-import from the TC database gets automatically updated when the database changes. As a collaborative tool, TC tries to utilise the effect of collaboration for the further standardisation of linguistic glosses. Glossing rules are conventional standards and one way to spread them is to make them easily accessible at the point where they are actively used. TC insists that annotators use a pre-defined set of glosses which are rooted in the Leipzig Glossing Rules [5], but have been extended to meet the needs of annotation in a more multi-lingual setting. We have grouped all glosses into annotation classes and mapped them to the GOLD (General Ontology for Linguistic Description) via URI-pointers. GOLD has been created as a tool to facilitate a more standardised use of basic grammatical features [3]. Through the linking to GOLD, TC users can relate annotations to grammatical concepts and to find relevant bibliographic resources as they annotate.

At present TC has 300 users, and the number is growing at a steady pace. We started 2012 with 13 000 annotated phrases from 97 mostly African languages. Since 2005 the TypeCraft system is under constant development and the planning of new development is based on the feed-back from system users and interested colleagues. Most of the active TypeCraft users are junior linguists and graduate students, mostly working on a less-resourced languages. TypeCraft is user driven. At present 54.2% of the data in the TC database remains private due to pending publication. However, also part of that data is partially represented on TC-wiki pages most of which were created by users of the system. A feature liked by many users is the possibility to join an annotation group. Through workshops and regular classes in linguistic text annotation new user groups can discover the tool. In addition, TypeCraft has proven to be a useful tool for e-learning for linguists. The TC wiki *Classroom* namespace contains several examples of classroom collaborations involving TC.

2.2 The TypeCraft Editor

After having imported a text into the TC Editor, which is easily accessed from the TC sites navigation bar (*New text*), the text is run through a simple, but efficient sentence splitter. The user can then select via mouse click one of the phrases to enter into the annotation mode. The system distinguishes between translational, functional and part-of-speech glosses. Properties can be assigned

[5] TC runs in Firefox, Chrome, Safari and Opera.

to linguistic phrases as well as to words or morphemes. Under Construction description the user can add notes and write down open questions which are stored together with the rest of the annotations in the XML-schema. The possibility to report and retrieve notes is an important feature of an annotation tool which is in particular needed at the beginning of an analysis, for collaborative annotation and when new layers of annotation are added to previous annotations.

The annotation interface is a table. Information is ordered horizontally (the phrase and the free translation) and vertically, so that words and morphs are aligned with their baseform and their part of speech, as well as their gloss. TC assumes that free class morphemes are annotated for meaning while closed class items receive a gloss. Phrases in TypeCraft are Url encoded data and thus can be exchanged with other web applications. For further processing TC data can be exported as XML.

The system prompts the user for the Lazy Annotation Mode.[6] Manual annotation is further added by making TC-tags accessible from the TCwiki navigation bar as well as by drop-down menu under annotation. Other important resources, such as Ethnologue [4] can be directly reached from the TC Editor.

Since TypeCraft uses Unicode, every script that the user can produce on the PC can be entered into the browser. As of recent, TypeCraft offers the transliteration of Mandarin Chinese and Telugu, a Dravidian language of India, to make data that has been entered into the database using the original script more widely accessible. The export of data to the main text editors is one of the services that TC offers. TC tokens can be exported to Microsoft Word, OpenOffice.org Writer and LaTeX, which allows easy integration of databased material into the user's favourite text editing software. Although annotating in TypeCraft is still time consuming it is the re-usability of data that pays off. Export can be selected from the text editing window or from the SEARCH interface.

2.3 TypeCraft Search

A TypeCraft search operates on phrases, which means that the result of a query is a phrase level representation. Each line (or block) of the search result represents a linguistic phrase. In our experience, search results consisting of lists of sentences can be evaluated more easily by humans than lines of concordances. Search results come in two flavours, either as lines of sentences which allow a first quick scan of the data, or as blocks of IGTs which give the linguist access to the sentence internal annotations. Using general browser functionality, search results can be easily scanned. TypeCraft allows for complex searches on several tiers from the search interface where words or morphemes queries can relatively freely be combined with a search for specific glosses or combinations of glosses co-occurring either in a phrase, or on a word or a morpheme. Visitors of TC can at present search freely in the 45% of the private data on TypeCraft that has been made available for general search.

[6] Lazy Annotation Mode (LAM) is a function that automatically enriches annotation tables with word related information already known to the database.

3 System Description

TypeCraft is a based on a remote server and a web-based client running locally in the user's browser. The various components of the system and their interactions are shown in Figure 1.

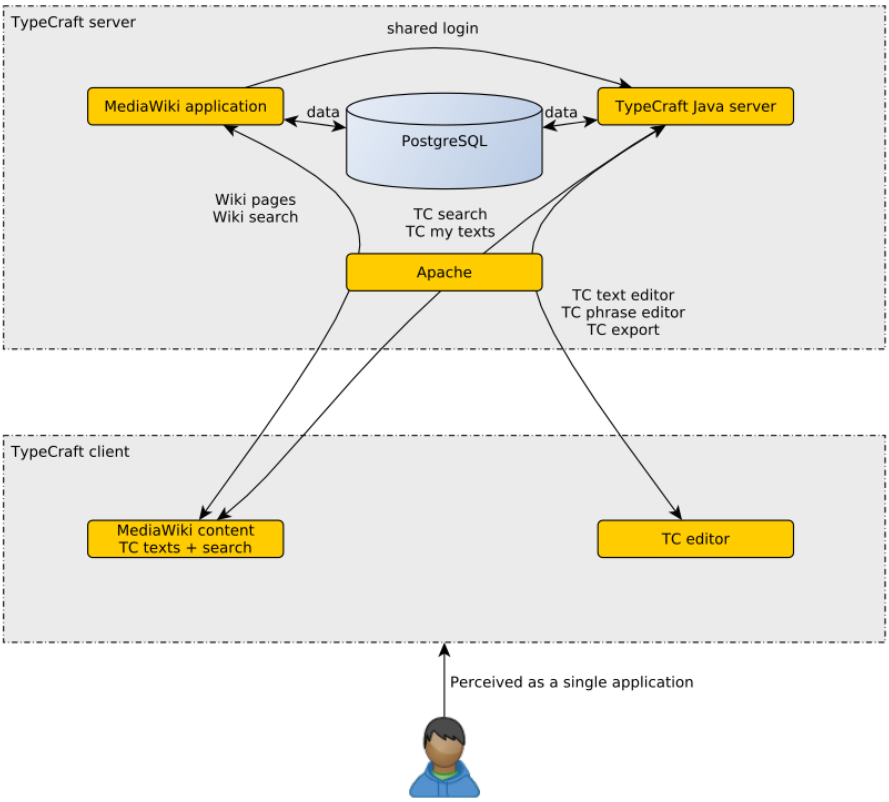

Fig. 1. TypeCraft architecture

3.1 Server

The TC server consists of a Java server application running in Tomcat, MediaWiki running in Apache, and a PostgreSQL database. The Apache server acts as a single entry point for both MediaWiki and the TC server application. MediaWiki has been extended with various extensions to enable specific TC functionality, such as *My texts*, *Phrase search* and *TC search*. The login is handled by MediaWiki but once authorised it will be shared with the TC server

application. This guarantees TC will only divulge data for which the logged in user has permission.

Storage and data model TC uses PostgreSQL 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. TC data can be divided into two specific groups:

- Common data: pos tags, gloss tags, global tags, ISO 639-3 languages. This is data shared between all annotated tokens and users.
- Individual data: texts, phrases, words and morphemes, together with their annotation. This is data specific to each user.

Individual data items reference common data items. E.g. everyone who uses the pos tag N will share the reference to a single common tag N.

3.2 Client

The user interacts with TC through a web-based interface. The interface consists of the customised MediaWiki content and a text-and-phrase editor (TC editor). The MediaWiki content provides wiki pages, wiki search, as well as the TC-specific functions mentioned previously. The editor is used to edit texts and phrases and assign annotation to phrases. It is written entirely in JavaScript and builds a GUI using HTML elements. The editor opens in separate browser window and is not directly connected to the MediaWiki content. The present GUI uses the YUI library and a large amount of TC-specific code. An important point is that for the end-user, TC is a single web application integrating the wiki and the editor.

4 Summary and Outlook

Setting it apart from the other main linguistic tools in its category, TC focuses on collaboration and the exchange of annotated linguistic data in the form of IGT. TC is low-tech on the outside. Non-computer-oriented linguists, language specialists and language communities can start using it almost instantly. While other linguistic software makes use of forums, blogs and other social software to service their user communities, TypeCraft *IS* social software. It is a powerful tool, however its real potential resides in its user community. Active users promote the creation and exchange of IGT, they link their data and contribute in this way to the availability, connectedness and quality of linguistic data.

One of the difficulties we faced over time was how to extend the data model when there are new requirements, or when a specific group needs a slightly different annotation model. Another issue is how to integrate TC annotations with third-party annotations on another level of the same data, e.g. audio and

video. We believe the right way to solve these issues is to switch to an RDF-based data model that will allow us to be flexible and also allow us to benefit from existing Linked Open Data (e.g. reference annotated tokens to WordNet). This will also open possibilities to create open RDF-based standards for linking heterogeneous annotations.

References

1. Beermann, D., Mihaylov, P.: e-Research for Linguists. In: Proc. of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (2011) 24–32
2. Broeder, D., Sloetjes, H., Trilsbeek, P., Van Uytvanck, D., Windhouwer, M., Wittenburg, P.: Evolving challenges in archiving and data infrastructures. In: Nau, H. N., Schnell, S., Wegener, C. (eds.): Documenting endangered languages: Achievements and perspectives (2011) 33–54
3. Farrar, S., Langendoen, T.: A linguistic ontology for the Semantic Web. GLOT International **7**(3) (2003) 97–100
4. Ethnologue: Languages of the World. `http://www.ethnologue.com/web.asp`
5. Leipzig Glossing Rules. Max Planck Gesellschaft (2010) `http://www.eva.mpg.de/lingua/resources/glossing-rules.php`