LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 4 (2010), pp. 60-65
http://nflrc.hawaii.edu/ldc/
http://hdl.handle.net/10125/4467

TypeCraft

from Pavel Mihaylov and Dorothee Beermann

Reviewed by Scott Farrar, *University of Washington*

1. OVERVIEW. The following review of TypeCraft is based on experience with the tool as of August, 2009. TypeCraft is an on-line database used to create, share, search, and present linguistically annotated texts (i.e., interlinear glossed texts). Users are able to create their own texts either by uploading files or by typing in texts by hand. TypeCraft aids the user in annotation by providing a visual framework for tokenization and by presenting hundreds of common linguistic categories. The data are sharable, since the tool is meant to encourage collaborative efforts: several people can edit a database at the same time. TypeCraft provides a rich search facility over the texts in the database, allowing users to search by linguistic level (phrase, word, morpheme) or by annotation element (gloss, feature, etc.). Finally, the tool can be used to present data in a variety of handy formats, including HTML, LaTeX, XML, and wiki markup.

2. OVERALL DESIGN. TypeCraft is designed in Java using current Web technologies, including a MediaWiki interface and a PostgreSQL database backend. The wiki interface provides a familiar, clean look and feel to the user experience. For instance, there is no need to have several programs open at once (editor, browser, terminal) while editing an annotated text. The wiki is easy to use, as the MediaWiki software (http://www.mediawiki.org) is quite familiar to many Web users (cf. http://wikipedia.org). Currently, TypeCraft runs only under the Firefox 3 browser.

2.1 ANNOTATION PROCESS. To begin the annotation process, the user is presented with a "Text editor" frame as shown in figure 1. The interface is similar to an on-line word processing application (such as Google docs). The user first chooses a language to assign to the text, and this is accomplished by using an auto-complete box. The list of languages is generated from Ethnologue (Lewis 2009; http://www.ethnologue.com). Once a language name is chosen, a link to the specific Ethnologue page for that language is created. For instance, if the language name "Spanish" is chosen, a link is created to:

http://www.ethnologue.com/show_language.asp?code=spa

The user is expected to use a language *name*, and not a language *code*, to search for the language entry. If the language name does not exist, then TypeCraft accepts the entry but throws an error when the user first tries to save the text. Thus, either the language has to be in Ethnologue or the string "undetermined" is used. Furthermore, the user can only use the primary language name (presumably generated from the open tables of Ethnologue).

The next step is to enter a text into the provided text box. Unicode is supported, though the user must enter the Unicode symbols into the text areas directly by cut and paste: there

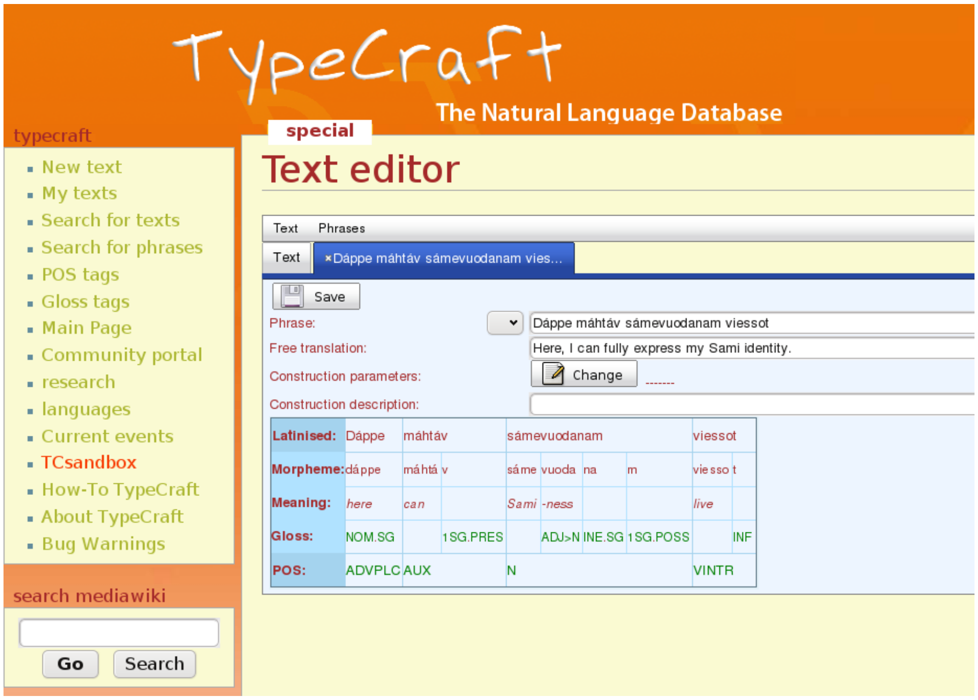
Licensed under Creative Commons
Attribution Non-Commercial Share Alike License

E-ISSN 1934-5275

is no soft keyboard built into TypeCraft. As a working example, consider the following Lule Sami text from TypeCraft's database:

(1) Gulluvasjvuohta ietjama duobddágijda sisñemusán vuona sinna la nanos. Dáppe máhtáv sámevuodanam viessot.

The user highlights each phrase to annotate, and TypeCraft builds a list of phrases, each of which shows up as a hyperlink in an adjacent area. If the user clicks on a link, TypeCraft then presents the text for annotation. The user can then begin to segment and insert delimiters (hyphens or spaces) between stems and affixes. Once this is done, TypeCraft builds a matrix for further annotation. As shown in figure 1, the matrix contains five rows: Latinised, Morpheme, Meaning, Gloss, and POS.



Text editor

Phrase: Dáppe máhtáv sámevuodanam viessot

Free translation: Here, I can fully express my Sami identity.

Construction parameters: -------

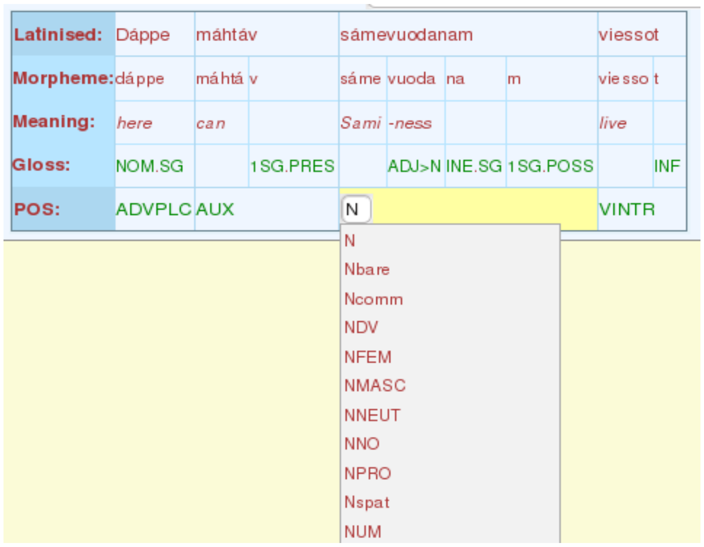
Construction description:

Latinised:	Dáppe	máhtáv		sámevuodanam				viessot	
Morpheme:	dáppe	máhtá	v	sáme	vuoda	na	m	viesso	t
Meaning:	*here*	*can*		*Sami*	*-ness*			*live*	
Gloss:	NOM.SG		1SG.PRES		ADJ>N	INE.SG	1SG.POSS		INF
POS:	ADVPLC	AUX		N				VINTR	

FIGURE 1: The matrix as built automatically from a text.

It is the task of the user, then, to tab through the matrix and enter various elements of annotation. For instance, when the user clicks on the POS (part of speech) box and begins to type, an auto-complete function is called which allows the user to search for the appropriate gloss element. The drop-down auto-complete box is shown in figure 2 for *N* (noun). TypeCraft also uses a "Lazy Annotation Mode," where glosses that have already been used are retrieved from the database and displayed to the user on subsequent annotations.

2.2 USE OF TERMINOLOGY. The terminology is stored internally within TypeCraft and cannot be customized. For instance, there are 66 POS category labels (e.g., PRON pronoun, REL relative, DET determiner) and 214 glossing tags (e.g., HUM human Animacy, ADD additive Aspect).



Latinised:	Dáppe	máhtáv		sámevuodanam				viessot	
Morpheme:	dáppe	máhtá	v	sáme	vuoda	na	m	viesso	t
Meaning:	*here*	*can*		*Sami*	*-ness*			*live*	
Gloss:	NOM.SG		1SG.PRES		ADJ>N	INE.SG	1SG.POSS		INF
POS:	ADVPLC	AUX		N				VINTR	

FIGURE 2: The auto-complete menu for POS.

These are the same categories used in TypeCraft's search facility (discussed below).

2.3 COLLABORATIVE ASPECTS. As for collaboration, the design allows for the single user (possibly with several different work spaces) as well as for groups of users. The latter adds an element of privacy to data collaboration, either because the project is not yet finished, or because of the cultural sensitivity of the material. Every time a text is created, the option of sharing with a particular group is presented. TypeCraft is well-suited for collaboration of native speakers with fieldworkers. The interface is fairly straightforward. In one scenario, the native speaker could enter the text and translation, while the linguist could add the annotation at a later stage. In another scenario, a group of linguists could use the tool, each responsible for annotating a particular aspect of the same text.

3. SEARCH FACILITY. A major appeal of the TypeCraft system is the ability to search both across whole texts (by title and language) and across any element of the text (phrases, words, or morphemes). That is, it is possible to search the TypeCraft database using the familiar "exact word" or "exact phrase" match method. In addition, one can search using "exact morpheme" and according to the elements of annotation. The search-on-annotation feature provides for very detailed searches. For instance, one could search using the following combination of features:

- phrases
- containing light verbs
- containing activity verbs
- topicalized information

That is, search for "phrases containing light verb and activity verbs where information is topicalized." Categories for phrase-level entities are auto-generated based on two different tagsets, currently "default" and "label conventions." The following is an example of the combinable features (using AND or OR) at the morpheme level:

- active
- 3SG
- AUX

That is, search for "auxiliaries in combination with active voice and third person, singular number." The same tagsets (for POS and for glossing elements) are used in the annotation editor.

4. INTEGRATION WITH OTHER TOOLS. TypeCraft uses a PostgreSQL database backend. All information entered into the system is stored remotely, an example of "cloud computing." Cloud computing refers to Web-based applications such that both data and application reside not on the user's machine, but "in the cloud," i.e., on a remote server or group of servers. Cloud computing frees the user from having to worry about data loss or software upgrades. Of course, the limits of cloud computing are not inconsiderable, especially for fieldworkers who may not have Internet access. As a consequence of all user data being stored remotely in a database, data projects are potentially affected by global database changes. For instance, if the database maintainers were to change an annotation tag from TNS to TENSE, then the change would show up in all records.

All data from the TypeCraft database may be exported in a variety of formats, including HTML, LaTeX, or XML. Furthermore, to encourage the sharing of data on wikis, texts can be exported in wiki format. Both the HTML and LaTeX formats are based on a simple tabular presentation of the data, while the XML output conforms to an XML schema containing nested elements (Phrase, Word, Morpheme, POS). And finally, since TypeCraft is designed as part of a wiki system, media including sound files and images are easily incorporated into the user's site. See, for example, this page on Èdó: http://www.typecraft.org/tc2wiki/User:Ota. It is currently not possible to upload Toolbox files into the TypeCraft database.

5. WEBSITE AND DOCUMENTATION. The TypeCraft site (http://www.typecraft.org) is organized using MediaWiki software. The site actually contains the application, its documentation, and the community portal, plus the wiki. As such, the wiki is a kind of one-stop interface for all of TypeCraft's functionality. There is a step-by-step tutorial that walks a user through a typical annotation session. The details of the tutorial are adequate and not overly detailed, as it is presented as a "quick start." To be considered as part of the documentation are the tables of linguistic categories (POS and glosses). For a given element of annotation, the tables list the name, abbreviation, and class to which the element belongs. Finally, there is a presentation of how TypeCraft data can be embedded into a wiki system, something that is quite useful considering the nature of TypeCraft itself.

6. COMPARISON WITH OTHER TOOLS. Another tool with similar functionality is the Field Linguist's Toolbox (available at http://www.sil.org/computing/toolbox). To compare TypeCraft with Toolbox, a number of dimensions can be considered, including platform, basic functionality, and interlinear text creation/annotation. The obvious difference of course is that while TypeCraft lives "in the cloud," Toolbox is a stand-alone desktop application to be downloaded and installed locally. As a result, TypeCraft is arguably better suited in terms of platform, as it can run on any machine with Firefox, including Windows, Mac, and most varieties of Linux. Toolbox runs only on Windows and Mac (with questionable runability on Linux systems through a Linux emulator on Windows/Mac). In terms of basic functionality, Toolbox is the more all-in-one tool for the linguist who needs the ability to manage entire field projects involving lexicons and interlinear texts. TypeCraft is not intended for creating lexicons (though the developers plan to include such functionality in the future). Thus, the main comparison to be made between these tools concerns the creation and annotation of texts.

Toolbox, like TypeCraft, offers the user a semi-automated means to segment the text. With Toolbox the user can use a lexicon that has already been compiled to aid in the segmentation. In TypeCraft, the user performs this task manually by inserting, for example, hyphens between morphemes. Unicode poses an issue for Toolbox users, because success is based on the user's local system and whether certain fonts are installed. With TypeCraft, Unicode support depends on the Firefox browser, a situation that is usually less problematic than a dependence on local fonts. Both applications allow for export to different file formats, such as XML. TypeCraft seems to offer an easier path to HTML and PDF (via LaTeX). In terms of glossing tags, TypeCraft comes with a predefined set, while Toolbox is largely free-form. This is perhaps not a meaningful comparison to the average single-project user, but it is a key aspect of TypeCraft that allows for a community of practice to form. That is, all data that are annotated using TypeCraft are searchable using a common interface. Linguists can see how others have marked up similar data, along with where and how particular tags are used. Arguably, the same could be said of Toolbox *if* a large enough database of Toolbox files were publicly available, something that does not currently seem to be the case.

7. CONCLUSIONS. TypeCraft, then, is a very promising on-line tool for the collaborative annotation of interlinear glossed texts. It allows for data creation, data sharing, and data presentation. In addition, TypeCraft provides a rich search interface for the data contained within the on-line database. Because it is implemented in the cloud, there is no need to install software (other than Firefox). And even for beginning students or for native speakers without linguistics expertise, the tool can be a useful way to create data. It has a one-stop interface implemented via the easy-to-learn Media Wiki software. The availability of hundreds of pre-defined linguistic categories is impressive.

This tool has some aspects that could be improved. The first is the inability to run on browsers other than Firefox (though it should be said that this situation is considerably better than with a tool that requires proprietary software or hard-to-install plug-ins). The search facility could be more user-friendly, in terms of both the slightly out-dated, pull-down-style menus and the allowable searches. Currently, any combination of features is allowed, and the choices are just too open ended. This could be viewed as a feature of the system, though for the average user, it is a little daunting. Though TypeCraft has an impressive inventory of linguistic categories, a more detailed explanation of each would be a convenient feature to add in the next version. Since TypeCraft is wiki-based, the explanations could easily be linked to a community discussion.

Primary function:	Linguistic annotation of interlinear texts; search over a database of texts; collaboration
Pros:	Collaborative computing in the cloud; one-stop interface for annotation and search; availability of many linguistic categories; several export formats (HTML, LaTeX, XML, wiki markup)
Cons:	Works only with Firefox; categories need explanation
Platforms:	Firefox 3
Open Source:	No
Proprietary:	Yes (but open for all to use)
Available from:	http://www.typecraft.org
Cost:	Free
Reviewed Version:	August 13, 2009
Application Size:	Unknown
Documentation:	http://www.typecraft.org/tc2wiki/Help:Contents

References

Lewis, M. Paul (ed.). 2009. *Ethnologue: Languages of the world*. 16th edn. Dallas, TX: SIL International. Online version: http://www.ethnologue.com.

Scott Farrar
farrar@uw.edu