TypeCraft : Language annotation in the context of African linguistics

Dorothee Beermann, NTNU, Trondheim, Norway

1 Introduction

In recent years the number of digital tools for language processing has increased rapidly. Huge databanks of linguistically annotated natural language text have been developed. Yet many of these efforts remain restricted to a few key languages. Although the development of language technology for African languages has increased drastically[1] in recent years, the usage and the maintenance of digital tools for language documentation is not as wide-spread as one would hope. There are many reasons, but one of the centre problems is that NLP tools standardly require expertise and access to resources that communities with minority or endangered languages lack. To make language preservation and documentation a general effort, ideally rooted in the local communities, we need to perceive a new generation of NLP tools with low technical threshold, but full functionality for language documentation and linguistic development.

Language documentation has many facets, and lexicography with the set-up of digital corpora and digital tools for dictionary development is one of its most central concerns. In the present paper we however will focus on a different facet of language documentation which we believe deserves more of our attention: At present language documentation is a linguistic field with very few ties to theoretical linguistics, yet it is evident that the availability of multi-lingual linguistically-processed data will eventually not only effect the way in which we think about language, but also have an impact on linguistic methodology and on the type of models we adopt to encode our linguistic understand. In short, we believe that the development of linguistic theory, and more general our perception of the nature of language, would benefit from the accessibility of language data also from lesser-known and lesser-documented languages. It therefore is a linguistic goal to make access and generation of linguistic-processed data a general commodity. Evidently different fields in linguistics have weighted the role of data and the form it takes in the presentation of linguistic research differently, but here is not the place to analyse these differences.[2] More central to our concern is that, as of now, studies of African and other Third-World languages still have a lower impact on the development of linguistic theory than studies

[1] For more information see: AfLaT (2008)

[2] For more information see Lehmann (2004) who explicitly discussed the role of interlinearized glossing in linguistic reseach and the standardization of glossing.

of the world's key languages. To achieve a wider and thus a more balanced use of linguistic resources, it is essential that more -lingual data becomes available to the research community as a whole.

Tying our two concerns together, namely availability of user-centered technology for language documentation and the availability of annotated linguistic data as a general commodity, we have developed a linguistic text annotator. The TypeCraft online database for natural language text, that we would like to present in this paper, has several novel features which make interlinerized morphological glossing easier.The annotation editor is connected to a relational database core. Next to presenting a user interface for word-to-word annotation for non-expert users, TypeCraft has wiki functionality which allows the sharing of data between system users, and the access to background knowledge just a mouse-click away.

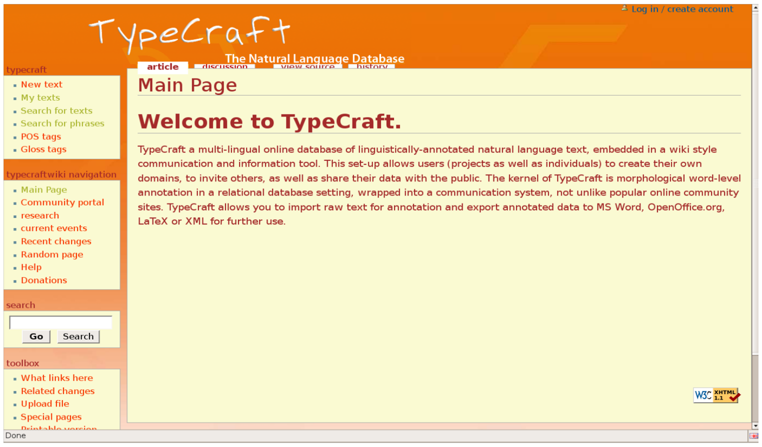
In this paper we would like to describe some of the main features of the TypeCraft application from the more comprehensive list presented below, and reflect their use in the context of African linguistics.[3]:

- import of text and individual phrases; automatic sentence splitting
- tablular annotation for word-by-word linguistic glossing with drop down reference guide for linguistic symbols and a menu for word or morpheme deletion or insertion
- knowledge sharing facility: sharing of data and background information between users of the system
- group features: keep your private space and share data with groups of your choice
- publish your data online using the TypeCraftwiki
- search the TypeCraft database

[3] A more detailed description of the system can be found in Beermann D. and Pavel Mihaylov: TypeCraft: Language Annotation for Human Beings. (in preparation)

- export of interlinear text to all main word processing programs suitable for use in linguistic papers
- export of xml for further automatic processing

TypeCraft a brief introduction



Main Page

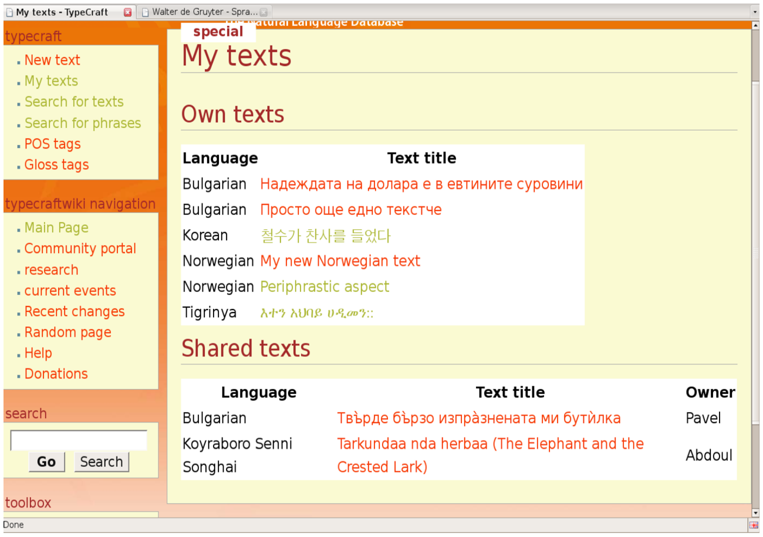
Welcome to TypeCraft.

TypeCraft a multi-lingual online database of linguistically-annotated natural language text, embedded in a wiki style communication and information tool. This set-up allows users (projects as well as individuals) to create their own domains, to invite others, as well as share their data with the public. The kernel of TypeCraft is morphological word-level annotation in a relational database setting, wrapped into a communication system, not unlike popular online community sites. TypeCraft allows you to import raw text for annotation and export annotated data to MS Word, OpenOffice.org, LaTeX or XML for further use.

Figure 1:

Typecraft is an online tool for interlinearized glossing of natural language phrases and small corpora[4]. Figure 1 displays the access point of TypeCraft through the **T**ype**C**raft wiki (TCwiki). Notice the navigation bar on the left in Figure 1; from here the user can (after login) enter *My Texts* which lets him view *Own texts*. In the case illustrated in Figure 2 the user also shares texts with other users, which appear under *Shared texts* together with information about the owners of these texts.

[4] At present TypeCraft can be used with Mozilla Firefox which can be freely downloaded from http://www.mozilla-europe.org/



My texts

Own texts

Language	Text title
Bulgarian	Надеждата на долара е в евтините суровини
Bulgarian	Просто още едно текстче
Korean	철수가 찬사를 들었다
Norwegian	My new Norwegian text
Norwegian	Periphrastic aspect
Tigrinya	እቶን አህባይ ሀዲመን::

Shared texts

Language	Text title	Owner
Bulgarian	Твъ̀рде бъ̀рзо изпрàзнената ми бутѝлка	Pavel
Koyraboro Senni Songhai	Tarkundaa nda herbaa (The Elephant and the Crested Lark)	Abdoul

Figure 2:

Different from Toolbox[5] which is a file-based mark-up system designed for linguistics data management, TypeCraft is a relational database, hosting annotated natural language text. Databases allow varies views of the data, and the possibility to search data using different search parameters. The TypeCraft database can for example be searched for *language, construction type, word, morpheme, annotation symbol.* Different from Toolbox, TypeCraft is an online system, and the user may search not only in his own data, but also in data that other system users have decided to share. Data can be viewed as part of texts, or by searching inside of individual tokens, which are phrases or sentences. TypeCraft uses unicode, so that every script the user can produce on his pc can be entered into the browser. Different from Toolbox TypeCraft insists on a set of linguistic tags, reflecting standards advocated for examples by the Leipzig Convention or initiative such a OLAC.[6] TypeCraft is a tool for any linguist interested in phrase or sentence annotation and linguistic data management, it is not an industriel tool meant for the annotation of huge corpora, however TypeCraft exports XML, allowing input to other applications. Before we turn to the description of annotating with TypeCraft some general remarks about the role of interlinerized glossing and of annotating linguistic data in general seems to be in order.

[5] Toolbox is a product of SIL International, an american organization originally committed to bible translation and the education of missionaries. SIL today is an organisation active in language documentation and the development of literary programs One of its main concerns are the documentation of the world's lesser-known languages. SIL can be found at: http://www.sil.org/

[6] OLAC can be found at: http:OLAC (2008)

Glossing

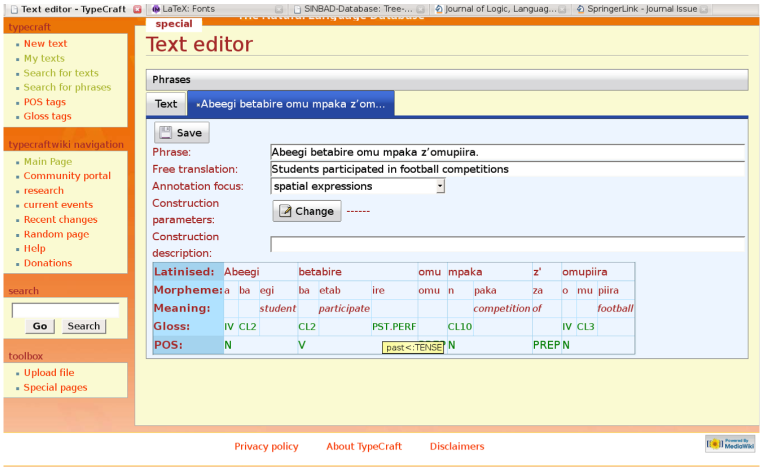
As pointed out by Lehmann (2004) the use of interlinearised glossing in the representation of primary data became a standard for linguistic publications as late as in the 80ies of the last century where glossing of sample sentences started to be required for all languages except those example sentences, coming from English. However, the use of glossed examples in written research was, and still is, not accompanied by a common understanding of its function, neither concerning its role in research papers nor its role in research. It seems that glosses, when occurring in publications, are seen by most linguistis as a convience to the reader, meant to facilitate the understanding of examples, especially from 'exotic' languages. This explains that data from a 'known' languages receive none or very little glossing. Moreover, information essential to the understanding of examples is often given in the article itself, and without any appropriate reflection in the glosses. As a result research rooted in different grammatical frameworks or geographically and politically distinct research communities, including research not published in English, will contain linguistic data that cannot be dechephired by all parties, and thus is lost to the linguistic community as a whole. But even with a clearer understanding of which role annotated natural language should play in linguistics, issues of standardization still need to be addressed. In this respect general linguistics seems to be lagging behind computational linguistics where standardization is vital to the development of the field. (TEI)[7].

Here we follow Lehmann (2004) and regard linguistic annotation, of which interlinerized morphological glossing is one specific form, as a linguistic method, and as such subject to rigour and coherence. Glossing that obeys standards allows the understanding of linguistic data independent of the research papers that presents it. However, to create interlinearized glosses as an independent linguistic resource requires that glossing becomes more standardized and more comprehensive. Yet, glossing is a costly process, which can only mean two things: already existing resources need to be made accessible, and the creation of new resources needs to be facilitated by the appropreate tools.

Glossing with TypeCraft

TypeCraft supports word-to-word glossing in an eight tier tabular setting as shown in Figure 3.

[7] For more information about natural language processing and standardization see for example: TEI (2008)



Text editor

Phrase: Abeegi betabire omu mpaka z'omupiira.

Free translation: Students participated in football competitions

Annotation focus: spatial expressions

Construction parameters: ------

Construction description:

Latinised:	Abeegi			betabire			omu	mpaka		z'	omupiira		
Morpheme:	a	ba	egi	ba	etab	ire	omu	n	paka	za	o	mu	piira
Meaning:			*student*		*participate*				*competition*	*of*			*football*
Gloss:	IV	CL2		CL2		PST.PERF		CL10			IV	CL3	
POS:	N			V				N		PREP	N		

Figure 3:

After having imported a text and run in through the sentence splitter, a process that we will not describe here, the user can click on one of the phrases to enter the annotation mode. Annotation is done in a tabular format where translational, functional and part-of-speech glosses are kept distinct. Every TypeCraft phrase, which can be either a linguistic phrase or a sentence is accompagnied by a free translation. In addition the specification of an annotation focus is complusary. The latter specification is made in additional input masks above the annotation table. Further meta-tags, identifying construction parameters, are possible. Under annotation the user has access to a drop-down menu, showing standard annotation symbols. In addition the TypeCraft wiki contains lists of (functional) glosses and pos symbols which provides further information about each symbol. In Figure 3 we also see the effect of mousing over symbols, which display the 'long-names' of the symbols. Some symbols belong to a metatag representing its supertype. In Figure 3 we see that *present* is a subtype of *tense*.Due to space limitation we cannot describe other useful features of the annotation interface, such as the representation of non-latin scripts, deletion and insertion of words and morpheme, the accessibity of several phrases under annotation, the syncronisation of annotation tokens with the text that they are part of, and many more.

Export of data

Export of data is one of the central functions of TypeCraft and export to some of the main editors (WORD and OPEN OFFICE) as well as Latex are possible. The user can export directly from his text editing window or from the SEARCH interface. Example (1) shows a sentence example exported from TypeCraft.

(1) *Runyankore-Rukiga, generated in TypeCraft*

Omu muti harimu enyonyi.

omu	mu	ti	ha	ri	mu	e	nyonyi
in.SPTL	CL3	*tree*	CL16	PRES	LOC	IV	*bird*.CL9
PREP	N		COP			N	

'In the tree there is a bird'

Example (1) is a sentence from Runyankore-Rukiga, a Bantu language spoken in Uganda. Example (1) illustrates locative inversion. The translational and function glosses, which belong to two distinct tiers in the TypeCraft annotation interface, reflecting the different nature of the information they encode, appear as one line when imported to one of the word processing programs supported by TypeCraft. Although glossing on several tiers is conceptual more appropriate, it does not comply with the editorial standards of linguistic papers. Some balence needs to be found, and as for now we have decided on an export displaying 5 tiers for languages with a latin script: in addition to the original string, one tier indicates morpheme bounderies, one translational and functional glosses, one part of speech tags, finally a free translation is given. Next to LATEX, TypeCraft examples can be exported as tables to Microsoft WORD and OPEN OFFICE. In addition xml export is possible.

TypeCraft and research on African languages

The TypeCraft annotation system originated as an annotation tool used by students studing language typology and multi-lingual approaches to construction typology as part of the linguistic graduate program at NTNU[8]. Some of the students working with the system were from Africa, and between the top ten languages documented in the database we find several Bantu languages, such as Runyankore-Rukiga and Swahili and West African languages, such as Akan and Sekpele. The statistics below reflects the number of tokens in the database at the end of Februar 2008. With at that time 646 tokens, that is, phrases or sentences, TypeCraft is yet a small database.

[8] NTNU is an acronym for the Norwegian University of Science and Technology: http://www.ntnu.no/english

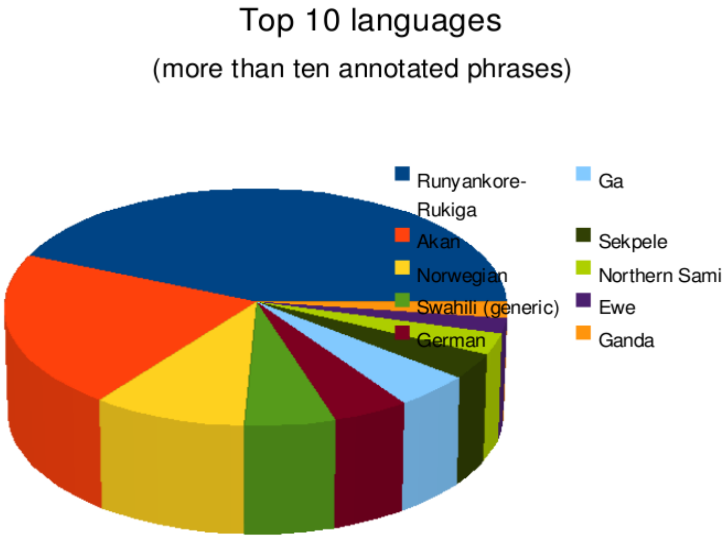


Table 1:

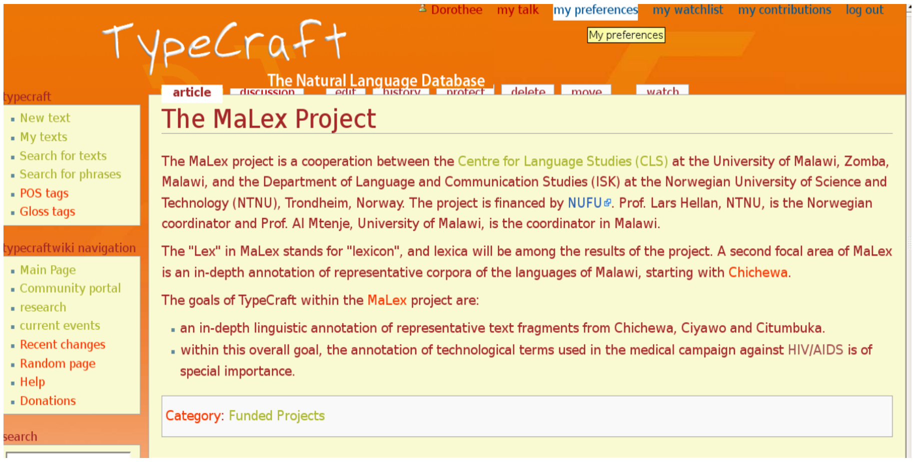
Of particular interest in the context of this paper is the MaLex project which uses TypeCraft to develop an in depth representation of some of the languages in Malawi. The project is a cooperation between the Centre for Language Studies (CLS) at the University of Malawi, Zomba and the Department of Language and Communication Studies (ISK) at NTNU. The "Lex" in MaLex stands for "lexicon", and lexica will be among the results of the project. A second focal area of MaLex is the annotation of representative corpora for some of Malawie's languages, starting with Chichewa. The MaLex project furthermore uses the TCwiki to share data and publish information about the project. Particular in context where the information about the goals and the progress of a language project is vital to its success, the wiki functionality of TypeCraft is of crucial importance. So let us look at the TCwiki in some more detail.

The TCwiki is an information sharing system, which is wrapped round the TypeCraft annotation tool. The wiki allows the TypeCraft users to collaboratively edit and organize the content of webpages. To this end we have customized a version of mediawiki [9] which is a software package best know from the wikipedia[10] The two important functions of the TCwiki are (i) the build-up of background knowledge essential for language annotation, and (ii) to provide a discussion formum for annotators and other TypeCraft users. Wikis in general have very powerful features; articles can be created and edited at anytime. Changes can be monitored, and the reversing of changes is possible. There are of course other versions of a content management systems, but mediawiki is

[9]

[10] Mediawki can be found at http://www.mediawiki.org/wiki/MediaWiki

probably the best known package for sharing information online. Distinct from mediawiki and special for the TCwiki is its domain feature. Every user manages his own interlinerized glosses. The MaLex project owns data which can only be viewed by members of the project. In addition the project can share text with individuals or other reserach groups which can be an informal research group or a funded project, so the system provides a safe development environment to the project. In addition interlinerized glosses can be shared directly with other TypeCraft users, if the wish arises. Figure 4 shows on of the pages representing the Malex project which will be instrumental in the promotion of the project within the local community and as an information platform for example relative to funding agencies.



The MaLex Project

The MaLex project is a cooperation between the Centre for Language Studies (CLS) at the University of Malawi, Zomba, Malawi, and the Department of Language and Communication Studies (ISK) at the Norwegian University of Science and Technology (NTNU), Trondheim, Norway. The project is financed by NUFU. Prof. Lars Hellan, NTNU, is the Norwegian coordinator and Prof. Al Mtenje, University of Malawi, is the coordinator in Malawi.

The "Lex" in MaLex stands for "lexicon", and lexica will be among the results of the project. A second focal area of MaLex is an in-depth annotation of representative corpora of the languages of Malawi, starting with Chichewa.

The goals of TypeCraft within the MaLex project are:

- an in-depth linguistic annotation of representative text fragments from Chichewa, Ciyawo and Citumbuka.
- within this overall goal, the annotation of technological terms used in the medical campaign against HIV/AIDS is of special importance.

Category: Funded Projects

Figure 4:

Conclusion

What we have described in this paper is the TypeCraft online tool for text annotation. TypeCraft is a database for natural language text, featuring a text annotator, accessible through a wiki which allows the sharing of information related to language annotation and documentation. TypeCraft can be found at www.typecraft.org. Since the TypeCraft database is in its initial state, we require a login. TypeCraft is designed for non-expert users. Its main goals are to make language data from lesser known languages available to a bigger research community, to add to the standardization of language annotation and to help projects, not least those in African linguistics, to promote their work.

References

AfLaT (2008). African language technology. http://www.aflat.org/.

Lehmann, C. (2004). Interlinear morphological glossingtowards a cognitive semantics. In Booij, G. E. C. L. J. M. S. S., editor, *Morphology Ein internationales Handbuch zur Flexion und Wortbildung*, pages –. deGryter Berlin-New York.

OLAC (2008). Open language archives community. http://www.language-archives.org/.

TEI (2008). Text encoding initiative. http://www.tei-c.org/index.xml.