# Interlinear glossing and its role in theoretical and descriptive studies of African and other lesser-documented languages

**Dorothee Beermann**
Norwegian University of Science and Technology
Trondheim, Norway
dorothee.beermann@hf.ntnu.no

**Pavel Mihaylov**
Ontotext,
Sofia, Bulgaria
pavel@ontotext.com

## Abstract

In a manuscript William Labov (1987) states that although linguistics is a field with a long historical tradition and with a high degree of consensus on basic categories, it experiences a fundamental devision concerning the role that quantitative methods should play as part of the research progress. Linguists differ in the role they assign to the use of natural language examples in linguistic research and in the publication of its results. In this paper we suggest that the general availability of richly annotated, multi-lingual data directly suited for scientific publications could have a positive impact on the way we think about language, and how we approach linguistics.We encourage the systematic generation of linguistic data beyond what emerges from fieldwork and other descriptive studies and introduce an online glossing tool for textual data annotation. We argue that the availability of such an online tool will facilitate the generation of in-depth annotated linguistic examples as part of linguistic research. This in turn will allow the build-up of linguistic resources which can be used independent of the research focus and of the theoretical framework applied. The tool we would like to present is a non-expert-user system designed in particular for the work with lesser documented languages. It has been used for the documentation of several African languages, and has served for two projects involving universities in Africa.

## 1 Introduction

The role that digital tools play in all fields of modern linguistics can not be underestimated. This is partially due to the success of computational linguistics and its involvement in fields such as lexicography, corpus linguistics and syntactic parsing, to just name some. Most crucially however this development is due to the success of IT in general and in particular to the World Wide Web which has created new standards also for linguistic research. Through the internet our perception of 'data' and publication of linguistic results has changed drastically only in a matter of a few years. Although the development of language resources and language technology for African languages is increasing steadily, the digital revolution and the resources and possibilities it offers to linguistics are mostly seized by researchers in the First World connected to work centering around the key languages. For this paper we would like to conceive of this situation in terms of lost opportunities: At present formal linguistics and linguistic research conducted on Third World languages are mostly undertaken with very little knowledge of each other and hardly any exchange of research results. Likewise, language documentation, which has roots in language typology and computational linguistics, only partially coincides with work in African linguistics. Yet, it is evident that the general availability of linguistic material from a bigger sample of languages will eventually not only affect the way in which we think about language, but also might have an impact on linguistic methodology and on the way we go about linguistic research. If you are only a few mouse clicks away from showing that a certain generalization only holds for a limited set of languages, but truly fails to describe a given phenomenon for a wider sample, statements claiming linguistic generality have to be phrased much more carefully. Our perception of the nature of language could truly benefit from general access to representative multilingual data. It therefore would seem a linguistic goal in itself to (a) work towards a more general

*Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages – AfLaT 2009*, pages 74–80,
Athens, Greece, 31 March 2009. 

and more straightforward access to linguistic resources, (b) encourage the systematic generation of linguistic data beyond what emerges from fieldwork and other descriptive studies and (c) advocate the generation of a multi-lingual data pool for linguistic research.

## 2 Annotation tools in linguistic research

It is well known that the generation of natural language examples enriched by linguistic information in the form of symbols is a time consuming enterprise quite independent of the form that the raw material has and the tools that were chosen. Equally well known are problems connected to the generation and storage of linguistic data in the form of standard document files or spread sheets (Bird and Simons 2003). Although it is generally agreed on that linguistic resources must be kept in a sustainable and portable format, it remains less clear, how a tool should look that would help the linguist to accomplish these goals. 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 not clear which direction research will take, which categories of data are needed and in which form the material should be organized and stored. But perhaps even more importantly most tools turn out to be so complex that the goal of mastering them becomes an issue in its own right. Researchers that work together with communities that speak an endangered or lesser documented language experience that digital tools used for language documentation can be technically too demanding.Training periods for annotators become necessary together with technical help and maintenance by experts which not necessarily are linguists themselves. In this way tool management develops into an issue in itself taking away resources from the original task at hand - the linguistic analysis.Linguists too often experience that some unlucky decision concerning technical tools gets data locked in systems which cannot be accessed anymore after a project, and the technical support coming along with it,has run out of funding.

### 2.1 TypeCraft an overview

In the following we would like to introduce a linguistic tool for text annotation called TypeCraft, which we have created through combining several well-understood tools of knowledge management. Needless to say, TypeCraft will not solve all the problems mentioned above, yet it has some new features that make data annotation an easier task while adding discernibility and general efficiency. That one can import example sentences directly into research papers is one of these features. In addition TypeCraft is a collaboration and knowledge sharing tool, and, combined with database functionality, it offers some of the most important functions we expect to see in a digital language documentation tool.

In the following we will address glossing and illustrate present day glossing standards with examples from Akan, a Kwa language spoken in Ghana, to then turn to a more detailed description of TypeCraft. However, a brief overview over the main features of TypeCraft seems in order at this point.

TypeCraft is a relational database for natural language text combined with a tabular text editor for interlinearized glossing, wrapped into a wiki which is used as a collaborative tool and for online publication. The system, which has at present 50 users and a repository of approximately 4000 annotated phrases, is still young. Table 1 gives a first overview of TypeCraft's main functionalities.

## 3 Glossing

The use of glosses in the representation of primary data became a standard for linguistic publications as late as in the 1980s (Lehmann, 2004) where interlinear glosses for sample sentences started to be required for all language examples except those 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 itself. It seems that glosses, when occurring in publications, are mostly seen as a convenience to the reader. Quite commonly information essential to the understanding of examples is given in surrounding prose, and often without any appropriate reflection in the glosses themselves.

Let us look at a couple of examples with interlinear glosses taken at random from the list of texts containing Akan examples. These examples are taken from the online database Odin at Fresno State University. The Odin database (http:/www.csufresno.edu/odin/) is a repository of interlinear glossed texts which have been extracted mainly from linguistic papers. The database it-

| Annotation | Collaboration | Data Migration |
|---|---|---|
| tabular interface for word level glossing - automatic sentence break-up | individual work spaces for users that would like to keep data private | manual import of text and individual sentence |
| drop down reference list of linguistic symbols | data sharing for predefined groups such a research collaborations | export of annotated sentence tokens (individual tokens or sets) to Microsoft Word, Open Office and LaTEX |
| word and morpheme deletion and insertion | data export from the TypeCraft database to the TypeCraft wiki | export of XML (embedded DTD) for further processing of data |
| lazy annotation mode (sentence parsing) | access to tag sets and help pages from the TypeCraft wiki | |
| customized sets of sentence level tags for the annotation of construction level properties | access to information laid out by other annotators or projects. | |

Table 1: Overview over TypeCraft Functionalities

self consists of a list of URLs ordered by language leading the user to the texts of interest.

### 3.1 The glossing of Akan - an example

Akan is one of the Kwa languages spoken in Ghana. The first example from the Odin database, here given as (1), comes from a paper by (Haspelmath, 2001)

(1) Ámá màà mè sìká.
Ama give 1SG money
'Ama gave me money. '

The second example is extracted from a paper by (Ameka, 2001):

(2) Ámá dè sìká nó máá mè.
Ama take money the give 1SG
'Ama gave me the money'

(Lit: 'Ame took money gave me')

The third example is quoted in a manuscript by (Wunderlich, 2003):

(3) ɔ-fɛmm me ne pɔflnkono.
3sg-lent 1sg 3sgP horse that
'He lent me a horse'

and the forth one comes from a manuscript by (Drubig, 2000) who writes about focus constructions:

(4) Hena na Ama rehwehwɛ?
who FOC Ama is-looking-for?
'Who is it that Ama is looking for?'

Except for Ameka, the authors quote Akan examples which are excerpted from the linguistic literature. Often examples coming from African languages have a long citation history and their validation is in most cases nearly impossible. When we compare (1) – (4) we notice a certain inconsistency for the annotation of *nó* which is glossed as 'the' (1), 'that' (3) and as DEF (2) respectively. This difference could indicate that Akan does not make a lexical distinction between definiteness and deixis, most likely however we simply observe a 'glossing figment'. The general lack of part of speech information in all examples easily leads us astray; should we for example assume that na in example (4) is a relative pronoun? The general lack of proper word level glossing makes the data for other linguists quite useless, in particular if they are not themselves native speakers or experts in exactly this language.*Màà* is a past form, but that tense marking is derived by suffixation is only indicated in (2) via a hyphen between the translational gloss and the PAST tag. Likewise *rehwehwɛ*in(4)is a progressive form, yet the lack of morpheme boundaries, and consistent annotation prevents that these and similarly glossed serve as a general linguistic resource. Purely translational glosses might be adequate for text strings which serve as mere illustrations; however, for linguistic data,that is those examples that are (a) either crucial for the evaluation of the theoretical developxment reported on, or (b) portray linguistic pattern of general interest, to provide morpho-

syntactic and morpho-functional as well as part of speech information would seem best practice.

It seems that linguists underestimate the role that glossing, if done properly,could play as part of linguist research. Symbolic rewriting and formal-grammar development are two distinct modes of linguistic research. Yet there is no verdict that forces us to express descriptive generalizations exclusively by evoking a formal apparatus of considerable depth. Instead given simplicity and parsimony of expression it might well be that symbolic rewriting serves better for some research purposes than theoretical modeling. One can not replace one by the other. Yet which form of linguistic rendering is the best in a given situation should be a matter of methodological choice. Essential is that we realize that we have a choice. Sizing the opportunity that lies in the application of symbolic rewriting, of which interlinear glossing is one form, could make us realize that the generation of true linguistic resources is not exclusively a matter best left to computational linguists.

## 4 A short description of TypeCraft

Typecraft is an interlinear 'glosser' designed for the annotation of natural language phrases and small corpora. The TypeCraft wiki serves as an access point to the TypeCraft database. We use standard wiki functionality to direct the TypeCraft user from the index page of the TypeCraft wiki to the TC interface of the database, called My Texts. My Texts is illustrated in Figure 1.The interface is taken from a user that not only possesses private data (Own texts), but who also shares data with other users (Shared Texts). At present sharing of text is a feature set by the database administrator, but in the near future the user will be able to choose from the TypeCraft user list the people with whom he wants to share his data. Note that data is stored as texts which consist of annotated tokens, standardly sentences. 'Text' in TypeCraft does not necessarily entail coherent text, but may also refer to any collection of individual tokens that the user has grouped together. A TypeCraft user can publish his data online; yet his own texts are by default 'private', that is, only he as the owner of the material can see the data and change it. To share data within the system or online is a function that can be selected by the user.

Different from Toolbox, which is a linguistic data management system known to many Africanists, TypeCraft is a relational database and therefore by nature has many advantages over file based systems like Toolbox. This concerns both, data integrity and data migration. In addition databases in general offer a greater flexibility for data search.For example, it is not only possible to extract all serial verb constructions for all (or some) languages known to TypeCraft, it is also possible to use the gloss index to find all serial verb constructions where a verb receives a marking specific to the second verb in an SVC. The other mayor difference between Toolbox and TypeCraft is that TypeCraft is an online system which brings many advantages, but also some disadvantages. An online database is a multi-user system, that is, many people can access the same data at the same time independent of were they physically are. Distributive applications are efficient tools for international research collaboration. TypeCraft is designed to allow data sharing and collaboration during the process of annotation. Yet although there are many advantages to an online tool, to be only online is at the same time a major disadvantage. Not all linguists work with a stable internet connection, and in particular for work in the field TypeCraft is not suitable.

TypeCraft uses Unicode, so that every script that the user can produce on his or her PC can be entered into the browser,[1] which for TypeCraft must be Mozilla Firefox. Different from Toolbox TypeCraft insists on a set of linguistic glosses, reflecting standards advocated for example by the Leipzig Convention distributed by the Max Planck Institute for Evaluationary Anthropology or an initiative such a GOLD (Farrar and Lewis, 2005).Yet, TypeCraft still allows a user-driven flexibility when it comes to the extension of the tag-set, as explained in the next section.

## 5 Glossing with TypeCraft

TypeCraft supports word-to-word glossing on eight tiers as shown in Figure 2. After having imported a text and run it through the sentence splitter, a process that we will not describe here, the user can select via mouse click one of the phrases and enter the annotation mode. The system prompts the user for the Lazy Annotation Mode (in Toolbox called sentence parsing) which will automatically insert (on a first choice ba-

[1] Note however that self-defined characters or characters that are not Unicode will also cause problems in TypeCraft

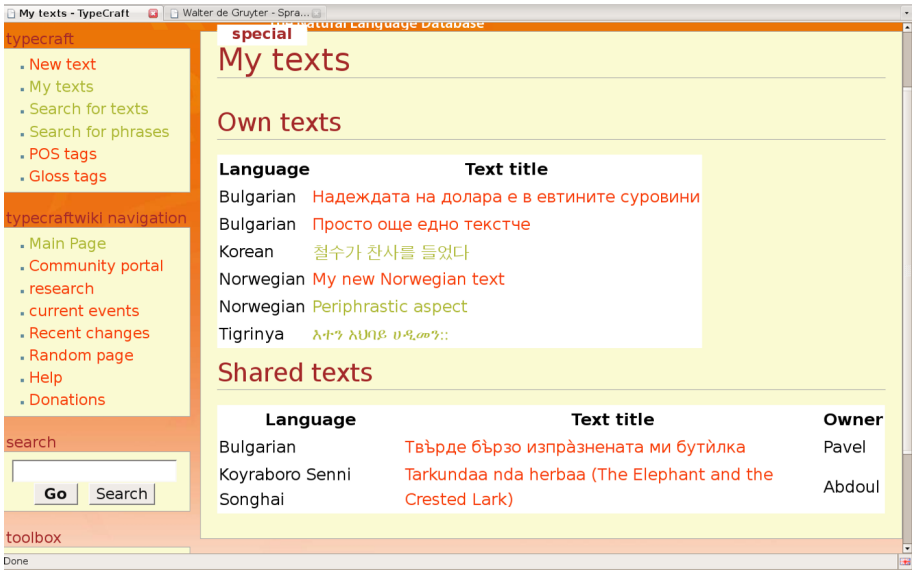


Figure 1: My texts in TypeCraft

sis) the annotation of already known words into the annotation table. TypeCraft distinguishes between translational, functional and part-of-speech glosses. They are visible to the annotator as distinct tiers called Meaning, Gloss and POS. Every TypeCraft phrase, which can be either a linguistic phrase or a sentence, is accompanied by a free translation. In addition the specification of construction parameters is possible. Although the user is restricted to a set of pre-defined tags, the TypeCraft glossery is negogiable. User discussion on the TCwiki, for example in the context of project work, or by individual users, has led to an extension of the TypeCraft tag set. Although TypeCraft endorses standardization, the system is user-driven. Glosses are often rooted in traditional grammatical terminology, which we would like to set in relation to modern linguistic terminology. The TCwiki is an adaquate forum to discuss these traditions and to arrive at a annotation standard which is supported by the users of the system. Under annotation the user has access a dropdown menu, showing standard annotation symbols. These symbols together with short explanations can also be accessed from the TypeCraft wiki so that they can be kept open in tabs during annotation. In Figure 2 we also see the effect of 'mousing over' symbols, which displays their 'long-names'. Some symbols have been ordered in classes. In Figure 2 we see for example that the feature past is a subtype of the feature Tense. This classification will in the future also inform search. Further features of the annotation interface that we cannot describe here are the easy representation of non-Latin scripts, deletion and insertion of words and morphemes during annotation, the accessibility of several phrases under annotation and the grouping of tokens into texts.

## 6 Data Migration

Export of data to the main text editors is one of the central functions of TypeCraft. TC tokens can be exported to Microsoft Word, OpenOffice.org Writer and LaTeX. This way the user can store his data in a database, and when the need arises, he can integrate it into his research papers. Although annotating in TypeCraft is time consuming, even in Lazy Annotation Mode, the resusablity of data stored in TypeCraft will on the long run pay off. Export can be selected from the text editing window or from the SEARCH interface.After import the examples can still be edited in case small ajustments are necessery.Example (5) is an example exported from TypeCraft.

(5)

| **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'*

(5) illustrates locative inversion in Runyakitara, a Bantu language spoken in Uganda. The translational and functional glosses, which belong to two distinct tiers in the TypeCraft annotation interface, appear as one line when imported to one of the word processing programs supported by Type-

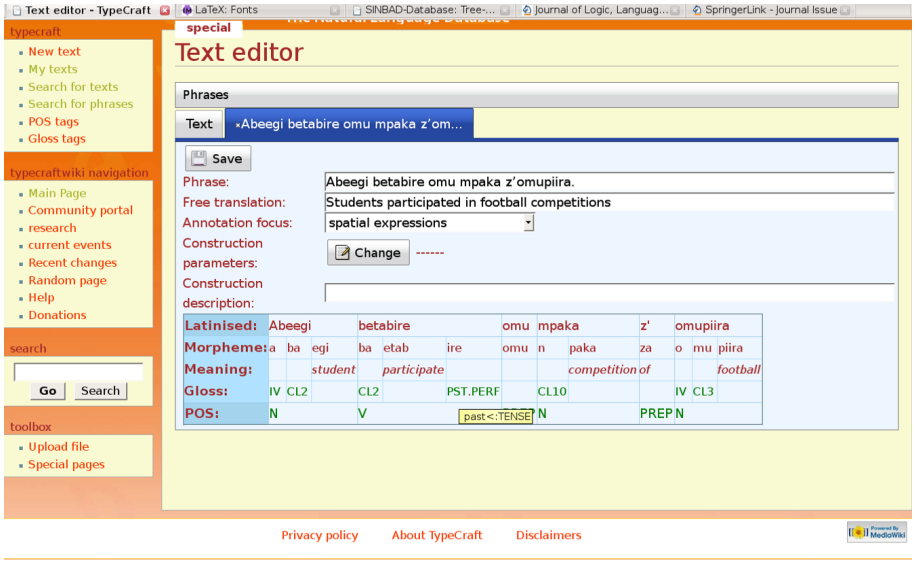


Figure 2: Glossing in TypeCrat

Craft. Although glossing on several tiers is conceptually more appropriate, linguistic publications require a more condensed format. As for now we have decided on an export which displays 6 tiers. Next to export to the main editors, TypeCraft allows XML export which allows the exchange of data with other applications. Figure 3 gives an overview over the top 15 languages in TypeCraft. In January 2009 Lule Sami with 2497 phrases and Runyakitara (Runyankore Rukiga)with 439 phrases were the top two languages. At present the database contains approximately 4000 from 30 languages.Most of the smaller language (with 300 t0 40 sentences) are African languages.

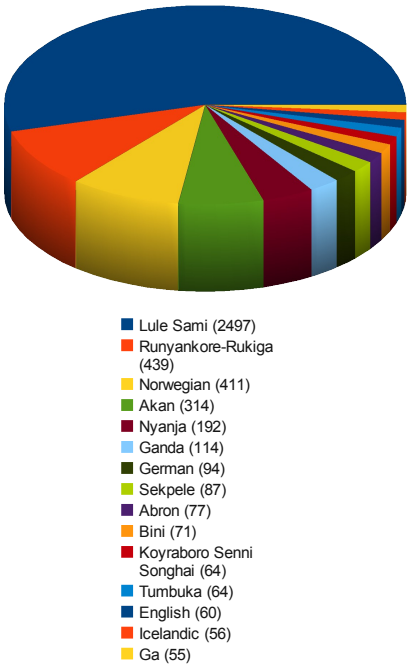


Figure 3: Top 15 TypeCraft languages by number of phrases

## 7 Conclusion

In this paper we suggest that the general availability of richly annotated, multi-lingual data directly suited for scientific publication could have a positive impact on the way we think about language, and how we approach linguistics. We stress the opportunity that lies in the application of symbolic rewriting, of which interlinear glossing is one form, and encourage the systematic generation of linguistic data beyond what emerges from fieldwork and other descriptive studies. With TypeCraft we introduce an online glossing tool for textual data which has two main goals (a) to allow linguists to gloss their data without having to learn how to install software and without having to undergo a long training period before the can use the tool and (b) to make linguistically annotated data available to a bigger research community. We hope that the use of this tool will add to the standardization of language annotation. We further hope that TypeCraft will be used as a forum for linguistic projects that draw attention to the lesser-studied languages of the World.

## References


Felix K. Ameka. 2001. Multiverb constructions in a west african areal typological perspective. In Dorothee Beermann and Lars Hellan, editors, *Online Proceedings of TROSS – Trondheim Summer School 2001*.

Hans Bernhard Drubig. 2000. Towards a typology of focus and focus constructions. In *Manuscript, University of Tübingen, Germany*.

Scott Farrar and William D. Lewis. 2005. The gold community of practice: An infrastructure for linguistic data on the web. In *In Proceedings of the EMELD 2005 Workshop on Digital Language Documentation: Linguistic Ontologies and Data Categories for Language Resources*.

Martin Haspelmath. 2001. Explaining the ditransitive person-role constraint: A usage-based approach. In *Manuscript Max-Planck-Institut für evolutionäre Anthropologie*.

William Labov. 1987. Some observations on the foundation of linguistics. In *Unpublished manuscript, University of Pennsylvania, USA*.

Christian Lehmann, 2004. *Morphologie: Ein Internationales Handbuch zur Flexion und Wortbildung*, chapter Interlinear morphological glossing. DeGryter Berlin-New York.

Dieter Wunderlich. 2003. Was geschieht mit dem dritten argument? In *Manuscript University of Düsseldorf, Germany*.