TypeCraft
Linguistic data and knowledge sharing
Open Access and linguistic methodology

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

Pavel Mihaylov
Ontotex
Sofia, Bulgaria
pavel@ontotext.com

Utrecht, June 15 -16, 2009
Small Tools for Cross-linguistic Research

...in the form of a handout:
This hand-out complements our demo of the TypeCraft (TC) system. We would like to raise questions concerning the design and the use of an application that combines online (and off-line) databasing and data annotation with a knowledge-sharing tool. Do we really need all this functionality, and if yes, how can we design a tool like TC in a way that makes it a useful device rather than just another gadget that consumes more of our time than it saves.

1 Background and Premises

Certain things we have to take for granted
(although we know we shouldn't...).
Let us briefly look at them.

1. One form of secondary linguistic data are interlinear glosses (IGs). In assuming that IGs reflect a first level of linguistic analysis we break with a tradition that sees them only as a convenience to the reader of linguistic publications. Good IGs are neither easy to find nor are they easy to make. Although we find glossed examples in the linguistic literature not all of them are in a format that would make it possible to use them in subsequent linguistic work. (Beermann, 2009). For one, glosses often lack the sufficient depth of

annotation to appear in work with a different research focus than the work that they were originally part of, and secondly, most of the examples found are already the result of long citation chains and the validity of the data is therefore hard to establish.
We would like to assume that linguistic research is dependent on reliable linguistic data which in turn means that the individual linguist has to either produce and/or have free access to such data.

2. Natural language resources are online (e.g., LDC[1]), yet they still might not be accessible to the individual linguist due to various circumstances, reaching from internal properties of the resource itself, such as its structure or ownership restrictions, to the fact that public interfaces to databases tend to be tedious to work with, to personal factors such as insufficient software support, or simply the lack of experience in dealing with digital resources. While the availability of online resource is not the same as the accessibility of these resources, by far not all existing resources become public, and at least in principle available.
Haspelmath (Haspelmath, 2009) mentions, for the field of language typology, the circumstance that databases cannot be published in any regular and accepted way as one of the potential reasons why valuable data all to often simply sits on a linguist's pc.
Although initiatives like DoBes offer online access to their archives, restrictions apply. Archiving at DoBes is subject to an approval process by the DoBes steering board and an advisory board. Clearly, archiving efforts need to observe different standards than online publication of digital databases. The latter must allow a more direct and easier access to the data than a language archive can permit. Archiving and publication of linguistic resources are two distinct processes subject to partially different sets of restrictions, and while archiving focuses on preservation, linguistic databases must focus on open access.

3. The Internet has created new standards for the accessibility of data and the need for new forms of publication.

4. We would like to assume a common understanding that standards for the annotation of linguistic material are needed for interoperability, without justifying this point any further here. Still, we

[1]Linguistic Data Consortium (Linguistic Data Consortium, 2009)

would like to refer to GOLD (General Ontology for Linguistic Description, 2009) and work by the Surrey Morphological Group (Surrey Morphological Group, 2009),contrasting two quotes to at least hint at the problem we cannot address here. Although standards are needed it remains an open question whether annotation standards can be derived from a 'set of universal linguistic features'. A feature is a theory dependent notion, and different annotation standards reflect the needs of a heterogenious linguistic community. Future standardization can only succed through an accommodation of different linguistic taxonomies.

> First, a common standard for the digitization of linguistic data may never be agreed upon; and the resulting variation in archiving practices and language representation would seriously inhibit data access, searching, and cross-linguistic comparison.
> *GOLD at www.linguistics-ontology.org/gold/2008*

> ...little is firmly established about features: we have no inventory of which features are found in the world's languages, no agreed account of how they operate across different components of language, no certainty on how they interact, and thus no general theory of features. They are used, but are little discussed and poorly understood. This is a central gap in the conceptual underpinning of much linguistic investigation.
> *Grammatical Features Home,*
> *Anna Kibort and Greville Corbett*

Instead of summarizing let us simply say that even if the linguistic world would be ideal (...at least from our perspective), that is, even if there would be a generally felt need for annotated data, accompanied by a keen interest in the standardization of descriptive linguistic categories, we still would have ways to go before we could base linguistic research on a shared linguistic methodology. In spite of interesting developements in linguistic databasing and online knowledge sharing, we neither have developed the right tools yet, nor can we offer at this point the infrastructure to support collaborative linguistic research in the age of electronic communication.

2 Linguistic methodology and digital linguistic tools

Let us now turn to the presentation of TypeCraft which is a linguistic application build around an online database. There are four main tasks that TypeCraft can help with. After we have given a demo of the system[2] we invite discussion concerning the functionality and the design of TC as well as some of its linguistic underpinnings.

TypeCraft is available at http://www.typecraft.org. At present a login is needed in order to use the database and the annotation editor. Certain pages of the TypeCraftwiki are public available. By the end of 2009 search as well as access to the wiki will be free, although annotation will still require a login.

The functionality of TypeCraft can be devided into four major tasks:

Task 1 generation of interlinear glosses

Task 2 managing data - locally and online; privat and in collaborations

Task 3 retrieval of data

Task 4 research cooperations and networking

We will demonstrate all four tasks, while the discussion will focus on glossing - word-level and meta-level and on the usefulness of database search.

3 Interlinear Glosses (IG)

1 to 5 below represent issues related to IG. 1 and 2 are discussed below while 3 to 5 are only mentioned and can be discussed if there is interest.

1. word level annotation and global annotation
2. why global annotation and how?
3. A TypeCraft text corresponds technically speaking to a list of tokens which are either individual sentences or sentences within a

[2] A short introduction to TC can be found at www.typecraft.org

running text. We intend to replace TC-texts by 'text' and 'collection' which will allow search in collections of sentences, for example representing locative inversion in Runyankitara, or text, for example about HIV/Aids in Chichewa.

4. We will introduce the OLAC standards for Metadata, the most important categories are illustrated below, and it seems as if they will serve our purpose well:

```
< xs:element name="title">
< xs:element name="creator''>
< xs:element name="subject">
<xs:element name="description"/>
<xs:element name="publisher"/>
<xs:element name="contributor"/>
<xs:element name="date"/>
<xs:element name="type"/>
<xs:element name="format"/>
<xs:element name="identifier"/>
<xs:element name="source"/>
<xs:element name="language"/>
<xs:element name="relation"/>
<xs:element name="coverage"/>
<xs:element name="rights"/>
```

5. validition of data (How good is an annotation?)
   Metadata systematically applied will help to judge the qualitity of the data.
   The TCwiki gives information about the annotators of TC data. Not all annotators are represented yet, and not all information that appears on their user pages seems equally useful for this purpose.

Ad 1: As illustrated in our presentation of the TC system, word level annoation happens on 4 tiers. We distinguish between part of speech annotation (POS tier) and annotation for functional and other properties of words and morphemes (GLOSS tier). A meaning tier holds the translational glosses. At present TC features 213 gloss tags and 17 gloss classes. The full overview over the gloss tags plus a brief description (which at present means spelling out of the symbols) and their gloss-class membership can be found in the TC wiki (accessed from the menu-

bar of the wiki). The wiki gloss tag page is automatically updated when the database changes. Comparing TC annotation with GOLD we find that GOLD distinguishes 11 categories under Morphosyntactic Properties. In some cases TC makes additional distinctions such as Aspect, Semantic Role or Deixis, in other cases we miss out on a distinction made in GOLD, such as Size. Often we use other categories to classify individual labels, we use for example the category Mood for the jussive while in GOLD it falls under Force.
Importantly, TC allows one or more symbols to represent the same descriptive category, such as ITER and ITR for *iterative* or M and MASC for *masculine*. Although we try to avoid one-to-many relations we nevertheless see the need to allow several symbols when these are established. We also allow different descriptive categories for what seems to be the same grammatical phenomenon. Grammatical categories are not only dependent on language specific grammar traditions but of course also theory dependent. So for example TC has a gloss OBJind for indirect object and a gloss DAT for the dative case. It it not hard to imagine that the first object in an English ditransitive construction might get annotated as DAT in spite of the lack of a case marker. In the generative tradition there are two candidates for so called dative constructions, namely the ditransitive construction and a construction where the subject receives dative case as illustrated by the German construction below:

positive-declarative-active-perception-copularVerb-predicative

Mir ist schlecht.

mir	ist	schlecht
1SG.DAT	*be*.PRS	*bad*
PRON	COP	ADJ

'I feel nausiated.'

'Indirect objects' in English may refer to the first object of a ditransitive construction or the prepositional object of a ditransitive construction, while in the German tradition only the dative object of a ditransitive construction is an indirect object, while a 'dative object' is the single object of a transitive verb marked for dative. Moreover in some approaches a 'case-marking preposition' is a notion so that the English preposition *to* for example could be glossed with DAT.
In a situation where one and the same descriptive category not necessarily refers to one and the same phenomenon while different cate-

gories might introduce an artificial distinction hiding a grammatical phenomenon cutting across languages, a mapping of descriptive categories which will allow the spelling out of known cross-categorizations would be desirable. In principle each descriptive category can be factorized allowing for a language + grammar tradition parameter. If such a representation were available it would be easy to identify deverging standards.

Ad 2: We us an 8bit system to tag global denpendencies related to the main predicate of a proposition. We distinguish the following categories: **Construction Kernel, Situation and Aspect, Voice and Frame Extensions, Additional Predicates, Theme-Rheme, Adjuncts, Illocution, Polarity**. Each of these categories has a list of subcategories. So for example the following list illustrates the subtypes of the Construction Kernel:

```
"auxiliaryVerb"
"modalVerb"
"intransitiveObliqueVerb"
"transitiveObliqueVerb"
"impersonalVerb"
"intransitiveRaisingVerb"
"intransitiveControlVerb"
"particleVerb"
"reflexiveVerb"
"transitiveRaisingVerb"
"transitiveControlVerb"
"lightVerb"
"inherComplVerb"
"transitiveVerb"
"idiom"
"intransVerb"
"multiple predicate kernel-coordination"
"multiple predicate kernel -SVC"
"transitiveParticleVerb"
"copularVerb predicative"
"copularVerb identity"
"ditransitiveVerb"
```

The Metatag system is under development. The following revisions of the present system are on the way: We will introduce Aspect as an in-

dependent category while Adjunct will not longer be on the list of construction parameters. In linguistics there is no consensus about which properties need to be present to linsence a certain construction label. Likewise, when several salient properties occur it is to a large extent up to the individual researcher which construction label (s)he chooses. Is the following construction from Runyankitare for example a passive or a causative construction or both?

Omucungwa kuganagisibwa omwana

òmùcùngwà	gùkànàgìsìbwà	òmwànà
o mu cungwa	gu ka nag is ibw a	o mu ana
IV CL1 *orange*	CL3 PST *throw* CAUS PASS IND	IV CL1 *child*
N	V	N

'An orange was made to be thrown by a child'

'Construction' and 'construction parameter' are conventional terms rather than scientific ones, yet metalabels making use of these terms have nevertheless an important function within TC: For one it is a way to open for database searches that target clusters of grammatical properties such as for all transitive verbs that express a situation of the type Emotion. Searches such as: "Give me all causatives, all negative propositions, all propositions with a secondary predicate", and many more now become possible.

4 Database search

TypeCraft is a small database. It is at present mainly used to generate sentence collections that illustrate certain linguistic phenomena; for example, datasets have been entered in conncetion with work on spatial expressions in Bantu. In this connection the TC wiki hosts an article preview on Locative Prepositions in Ruyankitara, a Bantu language spoken in Uganda.
Doctoral and masters work has added to the data TC hosts, especially for the Kwa languages of West Africa and on the Niger-Kongo languages of Uganda.

To get the feel for different kinds of data search we have compared search in a tagged corpus of Norwegian with search in Norwegian data in TypeCraft. We should bear in mind however that data extracted from a multi-lingual database of interlinear glossed sentences is of a different

nature than the data extracted from a corpus tagged mainly for part of speech categories. We expect that for many purposes a combined search of different resources will make the most sense.
The goal of any search is pattern recognition and classification, which for most linguists means an exercise in 'paper-and-pencil mental data-mining'.
For our little experiment in pattern visualization we searched for the Norwegian preposition *på*. Its base mining is 'on' and we were interested in uses of *på* that diverged from its base-meaning. Searching for *på* in TC resulted in a list of 34 sentences. Data visualization can be done nicely in TC by exporting the data set as an html document which allows easy browsing and a clean representation of the data. Export of patterns that are of particular interest to the user's favorite editor is possible, and has been illustrated in the presentation.
Our search in TC allowed us to isolate three main pattern, namely grammaticalized combinations between the preposition and certain verbs leading to new compositional meanings, idiomatisation and finally a productive pattern showing that *på* can occur in complementary distribution with the preposition *i*. The latter pattern is illustrated on one of our slides, but here is not the place to discuss it.
Pattern classification through search is of particular interest when comprehensive annotation reveals seemingly inconsistened annotations, as has happened in TC for the annotation of the Akan verbs corresponding to the English verbs *come* and *go*. At present the database has 4 active annotators working on Akan and its dialects. In order to make the information about the emerging distinct grammatical functions of these two verbs accessible to all annotators, we used the database search function and then data export to the TCwiki to document emerging patterns, and asked all annotators to help isolating the functions in question. The page is illustrated on one of your slides.

5 Research cooperations and networking

TC is at present used by two projects for data collection and project documentation. Our slides present the Malex project with is a collaboration between the University of Malawi and the Norwegian University of Science and Technology in Trondheim. The other project was a pilot study which had as its goal the creation and partial annotation of a small corpus of Lule Saami texts. Lule Saami is one of the endangered languages

of Europe. The Lule Saami project took place in 2008, and its results can be found on TypeCraft.

6 Some practical thoughts and some loose ends

We have presented in our demo our first attempt to facilitate annotation through LAM - the Lazy Annotation Mode. In particular for a possible continuation of the Lule Saami project we aim for the modular integration of a Saami morphological parsers into TypeCraft.

...

References

Beermann, D. (2009). From interlinearized glossing to standard annotation. In *Creating Infrastructure for Canonical Typology, Surrey*.

General Ontology for Linguistic Description (2009). http://linguistlist.org/emeld/gold-ns/index.cfm.

Haspelmath, M. (2009). Some thoughts on nyu's syntactic structures of the world's languages (sswel). In *online manuscript, http://unjobs.org/authors/martin-haspelmath*.

Linguistic Data Consortium (2009). http://www.ldc.upenn.edu/.

Surrey Morphological Group (2009). http://www.surrey.ac.uk/lis/smg/.