e-Research for Linguists

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

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

Abstract

e-Research explores the possibilities offered by ICT for science and technology. Its goal is to allow a better access to computing power, data and library resources. In essence e-Research is all about cyberstructure and being connected in ways that might change how we perceive scientific creation. The present work advocates open access to scientific data for linguists and language experts working within the Humanities. By describing the modules of an online application, we would like to outline how a linguistic tool can help the linguist. Work with data, from its creation to its integration into a publication is not rarely perceived as a chore. Given the right tools however, it can become a meaningful part of the linguistic investigation. The standard format for linguistic data in the Humanities is Interlinear Glosses. As such they represent a valuable resource even though linguists tend to disagree about the role and the methods by which data should influence linguistic exploration (Lehmann, 2004). In describing the components of our system we focus on the potential that this tool holds for real-time data-sharing and continuous dissemination of research results throughout the life-cycle of a linguistic project.

1 Introduction

Within linguistics the management of research data has become of increasing interest. This is partially due to the growing number of linguists that feel committed to the documentation and preservation of endangered and minority languages (Rice, 1994). Modern approaches to Language Description and Documentation are not possible without the technology that allows the creation, retrieval and storage of diverse data types. A field whose main aim is to provide a **comprehensive** record of language constructions and rules (Himmelmann, 1998) is crucially dependent on software that supports the effort. Talking to the language documentation community Bird (2009) lists as some of the immediate tasks that linguists need help with; interlinearization of text, validation issues and, what he calls, the handling of uncertain data. In fact, computers always have played an important role in linguistic research. Starting out as machines that were able to increase the efficiency of text and data management, they have become tools that allow linguists to pursue research in ways that were not previously possible.[1] Given an increased interest in work with naturally occurring language, a new generation of search engines for online corpora have appeared with more features that facilitate a linguistic analysis (Biemann et al., 2004). The creation of annotated corpora from private data collections, is however, still mainly seen as a task that is only relevant to smaller groups of linguists and anthropologists engaged in Field Work. Shoebox/Toolbox is probably the oldest software especially designed for this user group.Together with the Fieldwork Language Explorer (FLEx), also devel-

[1] We would like to cite Tognini-Bonelli (2001) who speaks for corpus linguistics and (Bird, 2009) who discusses Natural Language Processing and its connection to the field of Language Documentation as sources describing this process.

Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 24–32,
Portland, OR, USA, 24 June 2011. ©2011 Association for Computational Linguistics

oped by SIL[2], and ELAN[3] which helps with multimedia annotation, this group of applications is probably the best known set of linguistic tools specialised in supporting Field Linguists.

A central task for linguistic field workers is the interlinearization of text which is needed for the systematisation of hand-written notes and transcripts of audio material. The other central concern of linguists working with small and endangered languages is the creation of lexica. FLEx therefore integrates a lexicon (a word component), and a grammar (a text interlinearization component).

The system that is described here, assists with the creation of interlinear glosses. However, the focus is on data exchange and data excavation. Data from the Humanities, including linguistic data, is time-consuming to produce. However, in spite of the effort, this data is often not particularly reusable. Standardly it exists exclusively as an example in a publication. Glosses tend to be elementary and relative to a specific research question. Some grammatical properties are annotated but others that are essential for the understanding of the examples in isolation might have been left out, or are only mentioned in the surrounding text. Source information is rarely provided.

The tool presented in this paper tries to facilitate the idea of creating re-usable data gathered from standard linguistic practices, including collections reflecting the researcher's intuition and her linguistic competence, as well as data derived from directed linguistic interviews and discussions with other linguists or native speakers resulting in sentence collection derived from hand-written notes or transcripts of recordings. Different from natural language processing tools and on a par with other linguistic tools our target user group is "non-technologically oriented linguists" (Schmidt, 2010) who tend to work with small, noisy data collections.

2 General system description

Our tool consists of a relational database combined with a tabular text editor for the manual creation of text annotations wrapped into a wiki which serves as a general entrance port and collaboration tool. The system is loaded in a browser. The customised wiki serves as an access point to the database. Using standard wiki functionality we direct the user to the database via *New text*, *My texts*, and *Text- or Phrase search*. *My texts* displays the user's repository of annotations called 'Texts'. The notion of Text does not only refer to coherent texts, but to any collection of individual phrases. *My texts*, the user's private space, is divided into two sections: *Own texts* and *Shared texts*. This reflects the graded access design of the system. Users administer their own data in their private space, but they can also make use of other users' shared data. In addition texts can be shared within groups of users.[4]

Interlinear Glosses can be loaded to the systems wiki where they can be displayed publically or printed out as part of a customized wiki page. As an additional feature the exported data automatically updates when the natural language database changes.

Comparing the present tool with other linguistic tools without a RDBMS in the background, it seems that the latter tools falter when it comes to data queries. Although both the present system and FLEx share some features, technically they are quite distinct. FLEx is a single-user desktop system with a well designed integration of interlinear glossing and dictionary creation facilities (Rogers, 2010), while the present system is an online application for the creation of interlinear glosses specialised in the exchange of interlinear glosses. The system not only 'moves data around' easily, its *Interlinear Glosser*, described in the following section, makes also data creation easier. The system tries to utilise the effect of collaboration between individual users and linguistic resource integration to support the further standardisation of linguistic data. Our tag sets for word and morpheme glossing are rooted in the Leipzig Glossing Rules, but have been extended and connected to ontological grammatical information. In addition we offer sentence level annotations.

Glossing rules are conventional standards and one way to spread them is (a) to make already existing

[2] SIL today stands for *International Partners in Language Development.*

[3] http://www.lat-mpi.eu/tools/elan/

[4] At present data sets can only be shared with one pre-defined group of users at the time.

standards easily accessible at the point where they are actively used and (b) to connect the people engaged in e-Research to create a community. Glossing standards as part of linguistic research must be pre- defined, yet remain negotiable. Scientific data in the Humanities is mainly used for qualitative analysis and has an inbuilt factor of uncertainty, that is, linguists compare, contrast and analyse data where where uncertainty about the relation between actual occurring formatives and grammatical concepts is part of the research process and needs to be accommodated also by annotation tools and when it comes to standardisation.

2.1 Interlinear Glossing Online

After having imported a text into the Editor which is easily accessed from the site's 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 and in such a way enter into the annotation mode. The editor's interface is shown in Figure 1.

The system is designed for annotation in a multilingual setting. The user starts annotating by choosing the language for the text that she has loaded to the system from an integrated ISO-language list. Many languages of Africa are known under different names and it therefore is useful to find a direct link to the web version of Ethnologue, a SIL International resource. Ethnologue can for example help with identifying alternative language names and offers useful pointers to SIL publications. The present system distinguishes between different levels of annotation. Free translational glosses, standard for all interlinear glosses, and what we call construction descriptions are sentence level annotations; so is Global Tagging. These global tags can be selected in the form of eight construction parameters

Construction kernel: transitiveVerb, reflexiveVerb, multiplePredicate, transitiveObliqueVerb,...

Situation: causation, intention, communication, emotional-experienced, ...

Frame alternation: passive, middle, reflexive, passive+applicative, ...

Secondary predicates: infinitivial, free gerund, resultative,...

Discourse function: topicalisation, presentationals, rightReordering,...

Modality: deontic, episthemic, optative, realis, irrealis,...

Force: declarative, hortative, imperative, ...

Polarity: positive, negative

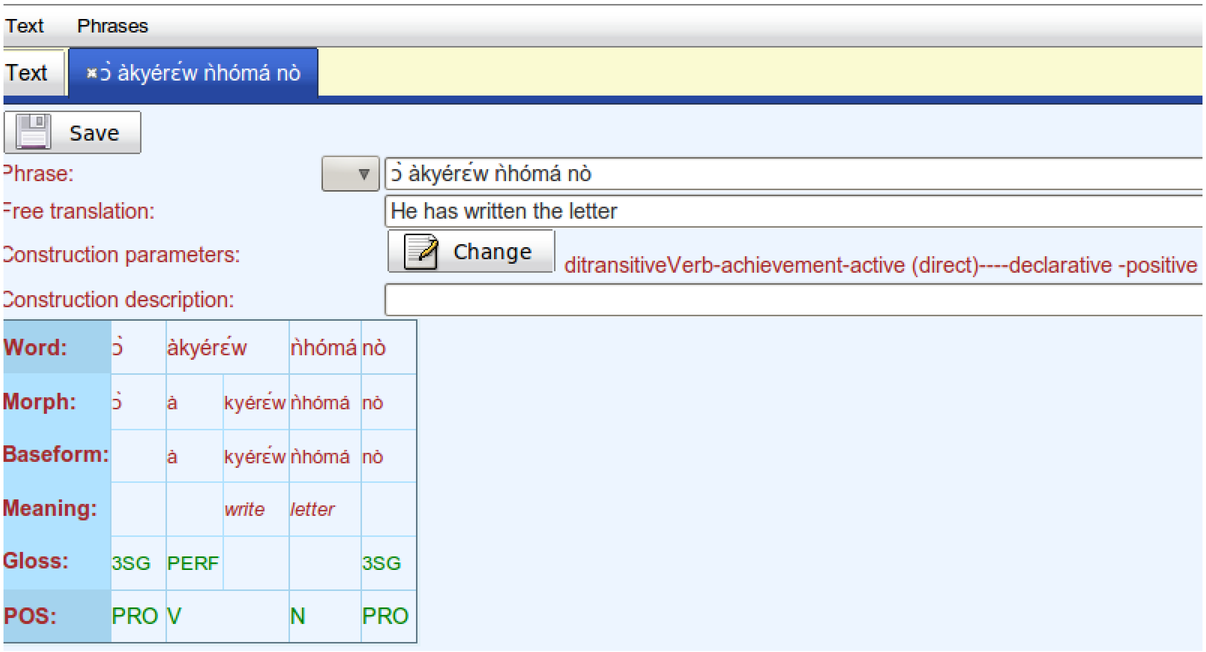
The field *Construction description* is meant for keeping notes, for example in those cases where the categorisation of grammatical units poses problems for the annotator. Meta data information is not entered using the Interlinear Glosser but the systems wiki where it is stored relative to texts. The texts can then fully or partially be loaded to the Interlinear Glosser. Using the wiki's Corpus namespace the user can import texts up to an individual size of 3500 words. We use an expandable Metadata template to prompt to user for the standard bibliographic information, as well as information about *Text type*, *Annotator* and *Contributor*. At present the corpus texts and the annotated data needs to be linked manually.

Word- and morpheme level annotation represents the centre piece of the annotation interface which appears as a simple table. Information is ordered horizontally and vertically, so that words and morphs are aligned vertically with their Baseform, Meaning, Gloss and Part of speech information. From the annotation table the user can chose one of the words and mark it as *Head* adding some basic syntactic information. Annotation can be partial and the idea is that free class morphemes are annotated for meaning while closed class items receive a gloss. Morphs may be accompanied by null to many glosses leading to enumerations of gloss symbols when necessary.

Each phrase has a unique identifier. This means that a data token can be shared freely online. The use case in Figure 2 illustrates this point.

Next to real-time data-sharing it is mainly the easy access to the relevant linguistic resources that facilitates manual annotation.[5]

[5] With the Lazy Annotation Mode (LAM) we offer an additional function that automatically enriches annotation tables



Phrase: ɔ̀ àkyérɛ́w ǹhómá nò

Free translation: He has written the letter

Construction parameters: ditransitiveVerb-achievement-active (direct)----declarative -positive

Construction description:

Word:	ɔ̀	àkyérɛ́w		ǹhómá	nò
Morph:	ɔ̀	à	kyérɛ́w	ǹhómá	nò
Baseform:		à	kyérɛ́w	ǹhómá	nò
Meaning:			*write*	*letter*	
Gloss:	3SG	PERF			3SG
POS:	PRO	V		N	PRO

Figure 1: The Interlinear Glosser

Three users of our system work together on the Bantu language Runyankore-Rukiga, a Bantu language spoken in Uganda. The language has no digital resources and annotated text is hard to come by. The group members experience a a lot of uncertainty in the selection of gloss values. While one of them is a lecturer at Makerere University in Kampala the other two study abroad. Mulogo attends class today, the topic is Tense and Aspect. He remembers that Ojore who tends to work at home has recently annotated his Field Work transcripts for his thesis on Tense and Aspect. Ojore happens to be online. Mulogo quickly asks Ojore if he could link him the two examples that he had mentioned the other day. They illustrated the co-occurrences of the immediate past and the perfective marker *-ire*. Ojore links him the tokens in Skype. Mulogo opens them in his browser and asks the teacher if he could project the examples after the break for some discussion. Meanwhile Ojore discovers that Dembe had in some contexts identified a morpheme that he has glossed as the immediate past as a present tense marker. Dembe is not online right now, so he links the two crucial examples to her in an e-mail. Normally they talk online in the morning when the connection to Kampala is better. He also adds a note to the construction description of the tokens for Mulogo and Dembe to read later.

Figure 2: A use case illustrating real-time data sharing

First of all lists over tags can be accessed from the wiki navigation bar where they are automatically updated when the database changes. The tag lists can be ordered either according to Gloss class or alphabetically. Short explanations for the glosses are provided. We have grouped all glosses into annotation classes and mapped them to the GOLD (General Ontology for Linguistic Description) ontology (See Figure 3). The idea behind Gold (Farrar and Langendoen, 2003) is to facilitate a more standardised use of basic grammatical features. As an OWL ontology it presents features in terms of categories and their relations. At this point the integration with GOLD is only light-weight and meant to give users of the system direct access to an ontology over grammatical types supplemented by bibliographic information and further examples showing the use of categories. This way essential information is made available at the point where it is needed. Uncertainty about the meaning of gloss can be reduced this way.

An important feature of the Interlinear Glosser is that it allows export of data to some of the main text editors - Microsoft Word, OpenOffice.org Writer and LaTeX. The example below illustrates an exported interlinear gloss. In addition to export from the Interlinear Glosser, individual or sets of interlinear glosses can be exported from the SEARCH interface which we will discuss in the next section. Offering a solution to the issue of wrapping (Bow et al., 2003), which arises for the representation of interlinear glosses for long sentences,[6] the system allows a clean representation of annotated sentences of any length. In general the alignment of morphemes and glosses (optionally indicated by a dotted line) forms the body of the interlinear gloss, while the original string and the free translation are wrapped independently

Omu nju hakataahamu abagyenyi

m	nj		hkthm					bgyngy		
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'

The example illustrates locative inversion in Ruyankore-Rukiga, a Bantu language spoken in Uganda. The translational and functional glosses, which belong to two distinct tiers in our editor, appear as one line when imported to a word-processor. Although glossing on several tiers is conceptually more appropriate, linguistic publications require a more condensed format.
Although to annotate manually is time consuming, it is the re-usability of the data that pays off. The ease with which already existing data can be exported from the system in order to be integrated into publications is one way to make this point.

In addition to export to Text Editors the system allows also from the graphical user interface the export of XML. The Akan sentence *àkyérɛw ǹhòmá nò* , meaning 'he has written the letter' (see Figure 1) is given as an XML structure in Figure 4. Notice that *Construction descriptions* and *Global tags* are exported together with the word- and morpheme annotations. Used for machine to machine communication, the XML rendering of interlinear glossses has interested the linguistic community (see for example (Bow et al., 2003)) as a means to find a generalised model for interlinear text.

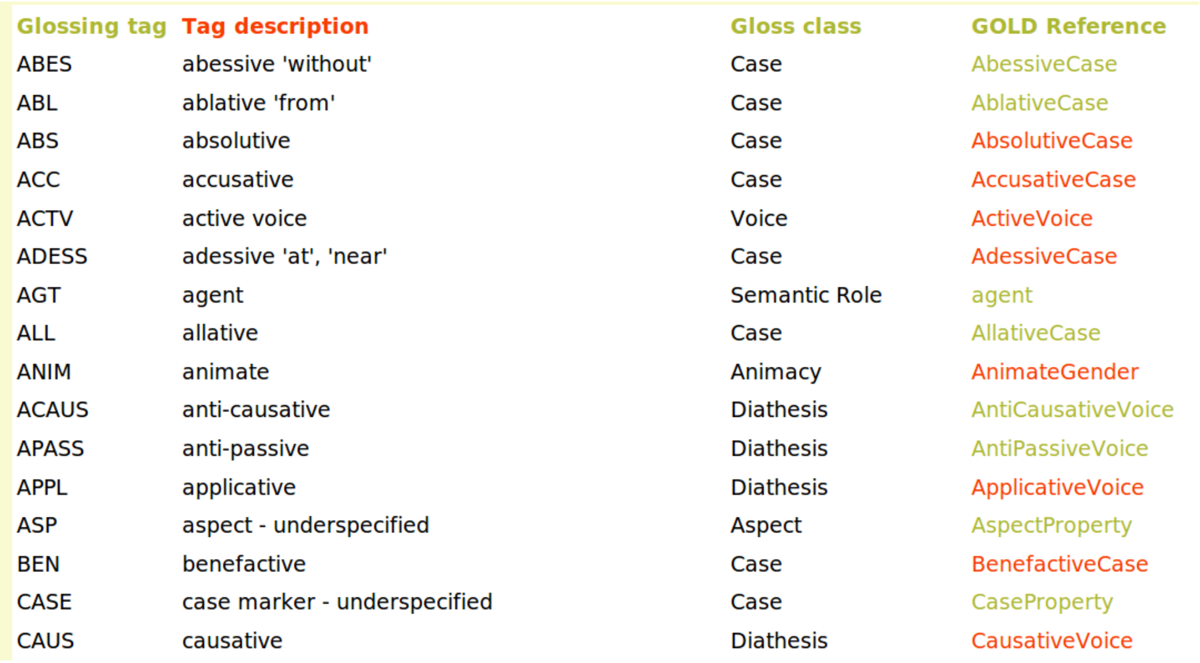
2.2 Search

Data queries operate on phrases, which means that the result of a query is a phrase level representation. Each line (or block) of the search result represents an individual sentence.[7] Lists of sentences, as the result of a search, are more easily evaluated by human observers than lines of concordances. Search results come as either lines of sentences which allow a first quick scan of the data or as blocks of interlinear glosses. This latter search output gives the linguist access to the sentence internal annotations. Using general browser functionality search results can easily be scanned. The system allows for complex searches from the graphical interface where word or morpheme queries can relatively freely be combined with a search for specific glosses or com-

with word related information already known to the database. LAM annotations need to be evaluated by the human annotator. They have only a limited value for languages with a rich system of allomorphic variation, but they are quite helpful otherwise even for languages with a rich portmanteau morphemes. In Toolbox this function is called 'sentence parsing'

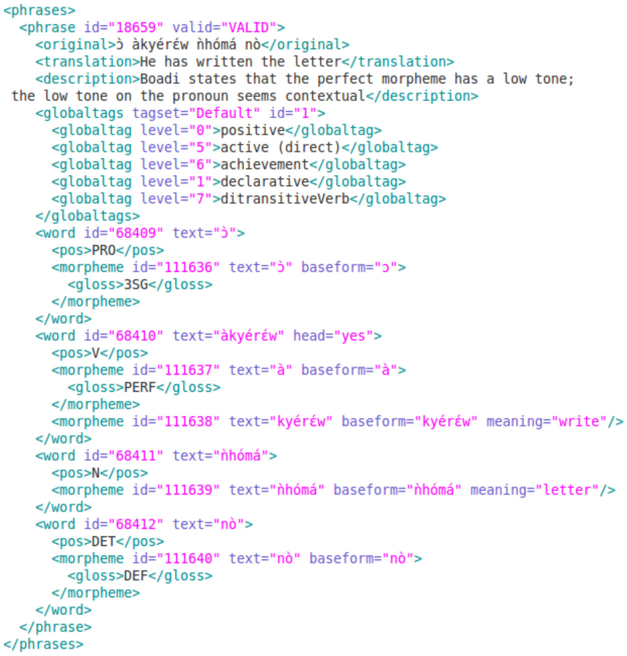
[6] What is a long sentence is a relative issue which is not only determined by the number of words that a sentence consists of, but also by the length of the enumeration of gloss tags that are aligned with each of the individual morphemes.

[7] or sentence fragment such as a noun phrase



Glossing tag	Tag description	Gloss class	GOLD Reference
ABES	abessive 'without'	Case	AbessiveCase
ABL	ablative 'from'	Case	AblativeCase
ABS	absolutive	Case	AbsolutiveCase
ACC	accusative	Case	AccusativeCase
ACTV	active voice	Voice	ActiveVoice
ADESS	adessive 'at', 'near'	Case	AdessiveCase
AGT	agent	Semantic Role	agent
ALL	allative	Case	AllativeCase
ANIM	animate	Animacy	AnimateGender
ACAUS	anti-causative	Diathesis	AntiCausativeVoice
APASS	anti-passive	Diathesis	AntiPassiveVoice
APPL	applicative	Diathesis	ApplicativeVoice
ASP	aspect - underspecified	Aspect	AspectProperty
BEN	benefactive	Case	BenefactiveCase
CASE	case marker - underspecified	Case	CaseProperty
CAUS	causative	Diathesis	CausativeVoice

Figure 3: Mapping between system-tags and GOLD concepts



```
<phrases>
  <phrase id="18659" valid="VALID">
    <original>ɔ̀ àkyérɛ́w ǹhómá nò</original>
    <translation>He has written the letter</translation>
    <description>Boadi states that the perfect morpheme has a low tone;
 the low tone on the pronoun seems contextual</description>
    <globaltags tagset="Default" id="1">
      <globaltag level="0">positive</globaltag>
      <globaltag level="5">active (direct)</globaltag>
      <globaltag level="6">achievement</globaltag>
      <globaltag level="1">declarative</globaltag>
      <globaltag level="7">ditransitiveVerb</globaltag>
    </globaltags>
    <word id="68409" text="ɔ̀">
      <pos>PRO</pos>
      <morpheme id="111636" text="ɔ̀" baseform="ɔ">
        <gloss>3SG</gloss>
      </morpheme>
    </word>
    <word id="68410" text="àkyérɛ́w" head="yes">
      <pos>V</pos>
      <morpheme id="111637" text="à" baseform="à">
        <gloss>PERF</gloss>
      </morpheme>
      <morpheme id="111638" text="kyérɛ́w" baseform="kyérɛ́w" meaning="write"/>
    </word>
    <word id="68411" text="ǹhómá">
      <pos>N</pos>
      <morpheme id="111639" text="ǹhómá" baseform="ǹhómá" meaning="letter"/>
    </word>
    <word id="68412" text="nò">
      <pos>DET</pos>
      <morpheme id="111640" text="nò" baseform="nò">
        <gloss>DEF</gloss>
      </morpheme>
    </word>
  </phrase>
</phrases>
```

Figure 4: XML export

binations of glosses. Search for portmanteau morphemes as well as for word-level co-occurrences of glosses is facilitated by allowing the user to determine the scope of gloss-co-occurrence which can either be the morph, the word or the phrase level. Queries are used to establish inter-annotator consistency, as well as to which degree an annotator is consistent in her annotations. For example, a search of 1154 Runyankore-Rukiga sentences, annotated by three different native-speakers in the context of different linguistic projects, shows that the annotators disagree on the meaning of the morpheme *-ire*. It is mainly annotated as PERF(ective) Aspect, but also as PAST, ANT(erior) and STAT(ive). However, when the same morpheme occurs in a negative context *-ire* is in 51 out of the 53 negative sentences annotated as expressing the perfective Aspect.[8] Although at present aggregate functions for the SQL queries can not be executed from the graphical user interface, the search offered by the system is already at this point a useful tool for linguistic data management.

3 Free data sharing and linguistic discovery

Collaborative databases where individual researchers or groups of researchers own portions of the data have their own dynamics and requirements for maintaining data sharing, recovery and integrity. They can be used with profit as an in-class tool or by research projects, and each of these uses requires a different set of rules for ensuring data quality and privacy. Annotations made by language specialists working on their own research reflect differences in interest and linguistic expertise.

Interesting data trends can be noticed by looking at the annotations made by annotators independently working on the same language. We will briefly illustrate this point with an example.

We have analysed the interlinear texts of four annotators working on individual linguistic projects in Akan, a Kwa language of Ghana. Together their work represents an annotated 3302 word corpus. We have analysed which glosses[9] were used and how frequently each of the glosses occurred. The most frequently used tags for Akan were SBJ and OBJ standing for subject and object, respectively. Comparing the Akan data with data coming from other users working on typologically distinct languages, we observe that the relative frequency in which the users annotate for the grammatical core relations ‘subject’ and ‘object’ differed from language to language.

As shown in Table 1 the absolute number of annotated morphemes and the relative frequency of SBJ and OBJ tags is highest for the two most configurational languages in our sample. This data has to be seen in the context of a possible use case not as the result of an empirical study. Other data tendencies indicative of annotator behaviour as much as of data properties can be observed too. Looking at Tense or Aspect within the same dataset shows that Akan which is a predominantly Aspect marking language (Boadi, 2008) (Osam, 2003) is by all four annotators mostly annotated for Aspect, with few tags for present tense. Between the Aspect tags we find HAB (habitual), as well as PRF and COMPL. The two latter glosses, referring to the *perfective* and the *completive* Aspect, where ‘completive Aspect’ means according to Bybee “to do something thoroughly and to completion”, might have been used to refer to a completed event. In the nominal domain it is the frequent use of the DEF gloss, as opposed to the very few uses of the gloss INDEF, that highlights that Akan marks definiteness but not indefiniteness. Interesting is that deixis is hardly marked although the definite marker in Akan has been claimed to have a deictic interpretation (Appiah Amfo, 2007).

The success of real-time data sharing depends on the trust that data consumers have in the data quality. All public data can be traced back to the annotator through the system’s Text search. As part of the first-time login procedure, each annotator is asked to contribute a small bio to her user page on the system’s wiki. In this way ‘data about the data’ is created and can be used to judge the data’s origin and authenticity. In addition an Advisory Board of senior linguists can be contacted for data review. Also, the list of Advisors can be viewed from the system’s wiki.

However, the kernel of all efforts is to assure that the data quality conforms to established criteria and procedures in the field. One way to accomplish this

[8] Date of query 03-03-2011

[9] The present survey does not cover pos tags.

Language	SUBJ	OBJ	units	SBJ %	OBJ %
German	5	2	1680	0,29	0,12
Norwegian	328	144	1787	18,35	8,05
Akan	470	393	4700	10	8,36
Kistaninya	0	0	737	0	0
R.-Rukiga	25	5	5073	0,50	0,10

Table 1: Relative frequency of core relational tags for 5 languages

is to link annotations to an ontology of grammatical concepts that reflects our present knowledge of grammatical categories and their relations. While we can work towards data validity, data completeness for a collaborative database will always depend on the linguistic goals pursued by the individual annotators.

It has been suggested by the GOLD community that the creation of Language profiles (Farrar and Lewis, 2005) could be a way to account for the morpho-syntactic categories of a specific language by using concepts found in GOLD under annotation. Given our own experience with the present integration of GOLD a mapping from the system's gloss sets to the GOLD ontology could be equally interesting. As an exercise in Ontology Extraction the mapping of annotation profiles from the present system to GOLD could as a first step allow the filling of category gaps. For the category CASE the equative is not yet known to GOLD, likewise Deixis and its forms such as proximate, distal, medial and remote are not currently represented.[10] It would be interesting to develop an algorithm which would allow to (a) build a model that can predict the 'class' of a certain gloss tag and (b) let ontological categories inform data search in the system presented here.

4 Conclusion

Data annotation and real-time data sharing requires a tool that is suitable for work in the Humanities. The system discussed here represents linguistically annotated data in the form of interlinear glosses, a well established format within philology and the structural and generative fields of linguistics. The present system is novel in that is allows the exchange of research data within linguistics proper. The systems's design has a clear focus on real-time data sharing combined with simplicity of use and familiarity of representation. It allows its users to concentrate on the linguistic task at hand. The system is particularly suitable for the creation of corpora of less documented languages.

While linguistic software makes use of forums, blogs and other social software, the present system *IS* social software. It is a powerful tool, however, its real potential resides in a growing user community and the effect that the community approach might have on data quality and the use of standards. Standards are ignored if not disseminated through an attractive public site that makes it easy for annotators to use them.With its relative longevity, and its institutional support, the system has two of the main characteristics of a digital tool that can serve as part of the cyberinfrastructure which is needed to support e-Research for the humanities (Nguyen and Shilton, 2008).

[10] Gold 2010 Data of search: 03/29/2011

References

Nana Appiah Amfo. 2007. Akan demonstratives. In Doris L. Payne and Jaime Pea, editors, *Selected Proceedings of the 37th Annual Conference on African Linguistics*.

Chris Biemann, Uwe Quasthoff, and Christian Wolff. 2004. Linguistic corpus search. In *Proceedings Fourth International Conference on Language Resources and Evaluation*, Lissabon.

Steven Bird. 2009. Natural language processing and linguistic fieldwork. *Computational Linguistics*, 35(3):469–474.

Lawrence A. Boadi. 2008. Tense, aspect and mood in Akan. In Felix K. Ameka and Mary Esther Kropp Dakubu, editors, *Tense and Aspect in Kwa Languages*. John Benjamins.

Cathy Bow, Baden Hughes, and Steven Bird. 2003. Towards a general model of interlinear text. In *Proceedings of EMELD Workshop 2003: Digitizing & Annotating Texts & Field Recordings. Electronic Metastructure for Endangered Language Data*. (EMELD) Project, May.

Scott Farrar and Terence Langendoen. 2003. A linguistic ontology for the semantic web. *GLOT International*, 7(3):97–100.

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

Nikolaus P. Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36.

Christian Lehmann. 2004. Data in linguistics. *The Linguistic Review*, 21(3-4):175–210.

Lilly Nguyen and Katie Shilton. 2008. Tools for Humanists. In D. Zorich, editor, *A Survey of Digital Humanities Centres in the United States*. Council on Library and Information Resources.

Emmanuel Kweku Osam. 2003. An Introduction to the Verbal and Multi-Verbal System of Akan. In Dorothee Beermann and Lars Hellan, editors, *Proceedings of the workshop on Multi-Verb Constructions Trondheim Summer School 2003*.

Keren. Rice. 1994. Language documentation: Whose ethics? In Lenore A. Grenobel and N. Louanna Furbee-Losee, editors, *Language Documentation: Practice and values*. John Benjamins.

Chris Rogers. 2010. Review of fieldworks language explorer (flex) 3.0. *Language Documentation & Conservation*, 04:78–84.

Thomas Schmidt. 2010. Linguistic tool development between community practices and technology standards. In *Proceedings of the LREC Workshop Language Resource and Language Technology Standards state of the art, emerging needs, and future developments*, Valletta, Malta, May. European Language Resources Association (ELRA).

Elena Tognini-Bonelli. 2001. *Corpus linguistics at work*. Amsterdam: Benjamins.