1

Glossing Language Online

DOROTHEE BEERMANN & ATLE PRANGE

Figure 1: Home of TypeCraft - www.typecraft.org

1 Introduction

Modern language documentation is among the essential tasks of linguistics. It redefines the borderlines between 'field linguistics', 'computational lin-

Texas Linguistics Society 10: Computational Linguistics for Less-Studied Languages.
Nicholas Gaylord, Stephen Hilderbrand, Heeyoung Lyu, Alexis Palmer and Elias Ponvert eds.

guistics', and formal linguistic research, and explores the potentials that lie in the combination of traditional field methods with new technologies. In the following we will discuss challenges and goals within language documentation and empirically based theoretical research. Our background is the development of educational programs within an international graduate program in linguistics at a Norwegian university and project-based research related to the languages of West Africa. In this paper we would like to present the project 'TypeCraft'[1] (TC) which focuses on the documentary and exploratory mode of research, and we in particular would like to highlight aspects of a comprehensive description of written language. We are interested in standards for interlinerized glossing of text and the classification of sentence strings as instantiating a 'construction type'. We will describe the functionality of the beta version of TypeCraft[2] (TC_{BETA}). The in-class use of TC and the experience we gained from testing TC_{BETA} on-line will be discussed. At present TC_{GAMMA} is under development, and we will briefly describe some of the basic design decisions we are facing.

2 Language Documentation: Challenges and Goals

The build-up of linguistically annotated language data faces two problems. The first problem concerns **data preservation**, including archiving, while the second problem resides in **data presentation**.

Outside of a larger project context, **forms of data preservation** are mostly dependent on the means chosen by the individual field linguist, and may vary greatly from the use of a tape recorder + on-paper documentation to digital techniques. Also within multi-lateral scientific projects, lexicographical work might consist of a series of individual book projects, collections of linguistic papers or the publication of dictionaries, while for other projects they may result in elaborated database systems for audio, video and written data. Among the computational tools for linguists, *The Field Linguist's Toolbox* (http://www.sil.org/computing/catalog/), distributed by SIL (the Summer Institute of Linguistics), should be mentioned.

---

[1] TypeCraft is a project financed by the faculty of Arts at NTNU, Trondheim and the Department of Language and Communication Studies at NTNU.

[2] Data documentation has been taught at the Department for Languages and Communication Studies of the Norwegian University of Science and Technology in Trondheim since 2004, where it is part of a seminar in language typology. One of the aspects that have been of particular interest from the beginning is the classification of language data according to construction types. We discovered soon that linguistically motivated categorization of natural languages is not an easy task, neither is the annotation of the individual sentences, taken seriously – so the name 'TypeCraft' was coined.

Yet, independent from how data is stored, one of the problems for the research community as a whole is that the material is often not accessible. The focus of the linguists conducting languages documentation is in most cases linguistic research, and the preparation of primary data is very work intensive. As pointed out by Berge (2004) in her review of the Kyoto Lecture on Endangered Languages, to publish results of fieldwork can take years. It in fact might look as if the idea that language data should be archived to assure public access potentially conflicts with the fact that most primary linguistic research is conducted as part of an academic career, which means that well annotated primary data is private until it is, as a book publication, made accessible. Public databases, on the other hand, maintained by universities or externally financed research project, like the 'Ailla' databank (http://www.ailla.utexas.org/site/welcome.html) for the Indigenous Languages of Latin America, seem to become more frequent, yet their existence is sometimes only known to a group of 'insiders'. In summary, raw or annotated primary language data is still sparse or simply hard to find. The necessity to archive natural language data does not seem as central an issue for the linguistic community as one would hope it to be.

In spite of efforts to standardize the annotation of language data from the side of computational linguistics as well as from within theoretical linguistics, **forms of written data presentation** are highly dependent on the individual linguist. Among attempts to standardize interlinarized glossing within theoretical linguistics is, e.g. the *Leipzig Glossing Convention* (http://www.eva.mpg.de/lingua/files/morpheme.html) together with the influential work by Christian Lehmann (2004), which, for example, is mentioned as a reference guide for contributors to the Folia Linguistica linguistics journal. Within computational linguistics the Gold Initiative under EMELD (http://emeld.org/gold-ns/index.cfm) presents an ontology of morphosyntactic terms, while the Text Encoding Initiative (TEI - http://etext.virginia.edu/standards/tei/teip4) represents a related initiative on a more general level of text encoding.

Still, as can be observed in theoretical writing, many unsolved problems remain. Open class items are in practice mostly glossed by corresponding words in English (or French), for example, without explicit indication of what the word's category is, with the consequence that possible categorical mismatches will not be detected. Closed class items are either treated like open class items, that is, they are simply translated into English, or analyzed, using sets of abbreviations that vary from language to language, from research tradition to research tradition, and from linguist to

linguist[3]. Evaluating the status of linguistic annotation from the use that is made of glossing in linguistic publications, one can safely say that we do not have a glossing convention, and primary data in theoretical linguistics papers do not satisfy the level of standardization needed for comprehensive language documentation. Still, we need to bear in mind that the use of interlinearized glosses is a relative new linguistic convention (for more discussion see Christian Lehmann op.cit.), and glosses in linguistic research are, as of now, rarely seen as the linguistic representations they are, but rather as a convenience to the reader. As a result, annotations of language samples within linguistic text are mostly idiosyncratic, while information that elucidates the example's functional and structural properties is given in the prose surrounding it. Moreover, what starts to emerge as an acceptable glossing standard for the well-studied languages within the Indo-European language family, and which has inspired automatic annotation tools like for example the 'Gold' standards, is not sufficient for most other languages, leading to several 'standards'. Within linguistic publications editorial standards have helped to introduce some glossing norms, yet, glossing itself is seldom understood as linguistic tool by itself. As a result language examples taken from linguistic literature can, in the majority of the cases, not be understood in isolation. A closer look at the publicly accessible Odin database will be illuminating in this context.

Summarizing: Language preservation and language presentation are concerned with

- making authentic linguistically annotated text available to the larger linguistic community.

---

[3] For instance, consider this example from Edo:

ù-khérhé-mwèn óghé ágá nà
nom-small-nom of chair this (continues next page)
`the smallness of this chair'

example#10 from the Odin language database (http://www.csufresno.edu/odin/

In the above example, the nominative seems to indicate that the category of the lexical item glossed as 'small' is that of a noun, yet we cannot be sure. 'Small' is treated as a definite expression in the English translation. What warrants this translation? The Edo expression ' ágá nà' seems to be a complex determiner, since the English translational gloss 'this' stretches over both words, yet the glossing might be misleading.

- archiving digitalized language data in a format that will survive to-be-expected system changes.
- promoting standards of annotation, particular within theoretical linguistics, so that the linguistic content of primary data can be understood independent of the particular research context that this data has been presented in.

3 How should a good glossing tool look?

According to our experience gained from lexicographical projects for West African languages, Shoebox/Toolbox is a management and analysis tool for field linguists that have many of the features needed to handle the linguistic annotation of text appropriately. It supports Unicode and can parse and interlinearize text. It offers the basic functionality of a linguistic tool, that is, it allows:

- import of text (manual, semi-automatic)
- parsing (morphophonemic)
- query (selection and projection)
- export (structured text)

Having taught Toolbox to masters students of linguistics and in the context of projects, we have observed that its use requires intensive training and constant guidance by an experienced user throughout a project. We furthermore observed that experienced linguists that were not familiar with automatic parsing, felt that data they had entered into Toolbox had 'disappeared', 'been altered' or could not be changed in the way they expected. With the beta version of TC we have developed a system that, although far from perfect, has a more intuitive user interface. This, we think, is one of the reasons why the mastery of TC can be attained in very short time. Next to the fact that a linguistic tool should not require a long training period to make professional use of it, another disadvantage of Toolbox is that it does not actively support annotation standards. Ideal in this respect would be a system that is designed such that it allows the user to focus on linguistic work and that it supports the standardization of glossing along transparent guidelines. As a step in this direction we built the TC_{BETA} glossing tool, as a web-based front-end to a relational database. The crucial advantage of an online system is that it allows direct matter oriented interaction between several linguists across time and space. Another advantage is that it can be used in teaching, since it makes the sharing of information easy, which fi-

nally, but not last, also means that it is a tool that supports the use of standard glossing conventions, since glossing is done in a public space.

4 TC$_{BETA}$ – aims and features

These were the three features that we believe define a good linguistic glossing tool:

- Its use must be intuitive – no training necessary before use!
- It must actively support encoding standards.
- It must safeguard the data and allow easy import and export to other systems!

4.1 A TC$_{BETA}$ Token

A record from TC. is a token identified for the language and the construction it represents, as shown in figure 2 below:

Sami : Locative deixis

Bievden li girje

Bievden		li	girje
bievdde	-n	li	girje
table	INE.SG	be.3PL.PRS	book.NOM.PL
N		COP	N

There are books on the table

Contribution by: Kristin Lindbach

TypeCraft token reference: 1158842886-Sami-Lule Sami-Kristin

Comments: In predicative constructions, the order of the NPs expressing 'location' and 'locatee'(the thing located) effects the interpretation of the 'locatee': If the 'locatee' follows the copula, its interpretation is indefinite (existential). (Cfr. possessive constructions: Máhtun li bednaga)

Figure 2: A TC-token - Lule-Sami '*There are books on the table*'

The string in bold in figure 2 corresponds to the script field in the annotation interface, while the first line in the table represents its Latin transliteration, which in Sami is identical to the script. The next line indicates morphological boundaries, followed by the two lines of glossing. The contributor line below the example is derived from one of the metadata fields in the annotation interface. Each token has a reference number; the annotators name and his comments accompany the token.

4.2 TC$_{BETA}$ – an online glossing interface

The editing interface of TC, displayed in figure 3, is divided into four main sections:

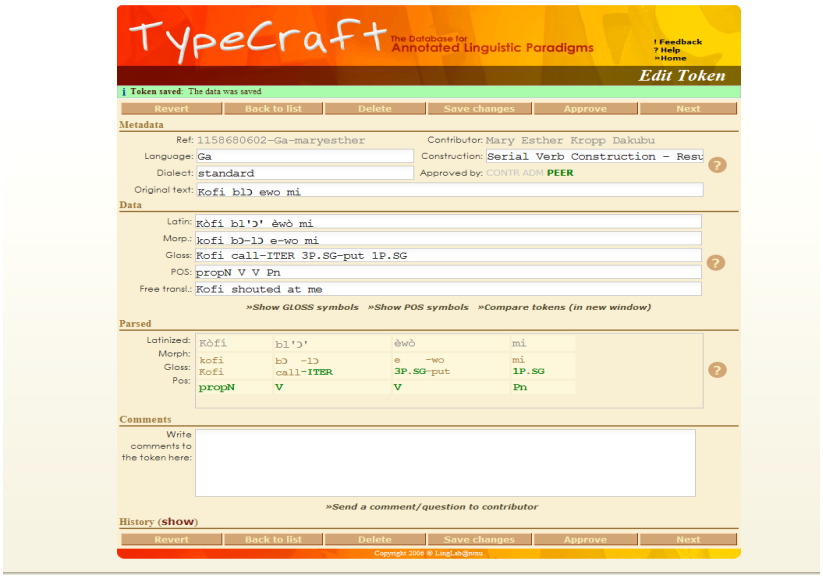

Figure 3 TC$_{BETA}$ -interface for data annotation

Glossing sentence strings is done in a 'single-window' approach. The horizontal divisions of the screen in four subsections display different types of data and different functionalities. The centerpiece of the interface allows the input of individual sentences. Input lines for original script and a Latin transliteration are provided. Using a layout that mimics the set-up of interlinerarized language examples known from linguistic research papers, the annotator assigns morpheme boundaries. Departing from the one-line linterlinarization used in linguistic publications, TC$_{BETA}$ supports two-line gloss-

ing, in the spirit of 'advanced glossing' (Drude 2002). As a first step to more conceptual clarity about the different type of information glossed, we at least would like to distinguish glossed categorical information from glosses for morpho-functional information. Underneath the 'editing field' is a 'parse field' where the annotator can follow the parsing of his input. In the parse field, translational glosses appear in a brown color, while recognized morpho-functional glosses appear in green. If a user enters a functional gloss that is not recognized by TC$_{BETA}$, the symbol will be parsed as a translation gloss, which means it will appear in brown rather than in green. In the line for categorical classes a symbol not known to the system will appear in red. In order to see which glosses are known to the system, the user can open a drop-down window in the same interface, which allows him to inspect available gloss symbols. If he is uncertain about the nature of the gloss symbol he can go to a HELP button on the same page, from where he can open an html file with more information about the use of glossing symbols in a new window. Different from other similar resources symbols are thematically grouped together and those symbols that are often used interchangeably are exposed. Moreover symbols representing properties of nominal inflection are discussed together, so are symbols sets used to annotate 'Aspect', etc. Notice, however, that nothing will prevent the user from integrating an unknown symbol into TC$_{BETA}$. Instead of being prevented from using unknown symbols, the user in encouraged to communicate his annotational choices. Two tools are offered at present:

I. The annotator is encouraged to make use of a *comment* field, where additional linguistic information can be communicated. If a user feels that, e.g., the first verb in a serial verb construction is neither a full verb nor an auxiliary and that one therefore should recognize a part of speech called *preverb* (PV), the comment field would be an obvious choice to enter this information. This way no information gets lost simply because existing standards have failed to make a distinction needed otherwise.
II. The annotator can take contact with TC$_{BETA}$ by pressing a button called FEEDBACK. At present this is a direct link to the administrator, but in the future this will be connected to other TC$_{BETA}$ users (see section 6). Here he can ask for support, not only oncerning the use of glossing symbols but any issue related to the annotation process.

In summary, the tools described above are first steps on the way to achieve standardization of glossing by offering on online site dedicated to the linguistic annotation of natural language data. The goal of TC_{BETA} has been to prove that annotation online is possible in the way we had conceptualized it, that is:

- To provide an online tool for the linguistic glossing of natural language
- To make documentation of glossing conventions and descriptions of construction types immediately accessible to the annotators.
- To foster glossing as a community effort, allowing the interactive use of annotation standards and their further development.

One of the next steps will be to allow the import of larger text for annotation, and a design that communicates the 'community spirit' of TC.

4.3 Desiderata for TC_{GAMMA}

We have tested TC_{BETA} in class room settings mainly with students from Africa, and at the University of Ghana in Legon; the system currently has 49 users. Two master students currently manage and annotate their data set with the help of TC_{BETA}.
Yet the system is tedious to use and at this point restricted to manual input through the browser; documents supporting the glossing process are still incomplete; and the flat list format of the display of tokens in the user's private domain is hard to use. Although users are in general comfortable with the annotation interface and the support of IPA and non Latinized scripts, the present lack of means to import small corpora and prior annotated data is a clear drawback. One likewise needs to find an easier way to export data to the standard editors and in xml format. Based on the users' communication with the administrator and our own observations, we compiled a list of desiderata for TC_{GAMMA}:

- TC must have a 'community profile' in the form of a multi-user topic related communication platform.

- TC must allow the import of small corpora and previously annotated data to represent data from less-studied languages, as well as less-known constructions from well-studied languages.

- TC must focus on the representation of multi-lingual construction types. Its profile is that of a typological database for richly annotated sentence strings representing one or more linguistic constructions.

- TC must be able to create language related word lists and help the user to retrieve inflectional and derivational paradigms connected to lexemes.

- TC must allow the user to employ one uniform data format, so that tokens are freely exportable from TC-online to main editors.

- TC must allow archiving; for example in the form of CDs or in national or international electronic repositories via an 'export button'.

5 The use of 'constructions' in Language documentation

The goal of any form of language documentation is that data be in some relevant sense representative for the language to be documented. Ways to achieve this goal are corpora, multi-media documents and lexica. The position we would like to defend here is that the character of a language is also present in the language's system of construction types. By this we mean that a language can be characterized by the set of subcategorisation frames and the array of valence alternations that it allows relative to these basic syntactico-semantic frames. We would like to call this inventory of constructions, when compiled over all open class items, the *Signature* of a language.

5.1 Construction layers

'Construction', as we use the notion here, is a descriptive term and should not be confused with the use of the term within Construction Grammar. A construction in the context of TC describes sentences with a certain set of syntactic and semantic properties. Any given sentence represents several constructions and working with a tool like TC, the user will have to decide which of the properties he would like to highlight by way of a certain construction label. A construction always represents a set of grammatical parameters that conspire, with some properties more central than others. Not only in active voice, but also in passive sentences, English verbs will agree with their subjects, still in order to illustrate subject-verb agreement, the use

of declaratives in active voice seem to be more suitable to exemplify subject-verb agreement than passive sentences or interrogatives. In a constructional network not all constructions have the same linguistic status. Subcategorisation frames represent what one might call the kernel of the constructional network. Frames derived by valence alternations constitute a further layer of constructions, while frames which alternate the information structure of the sentence provide a further layer and so does spatial-temporal modification. Figure 4 below gives a graphical summary of this point:

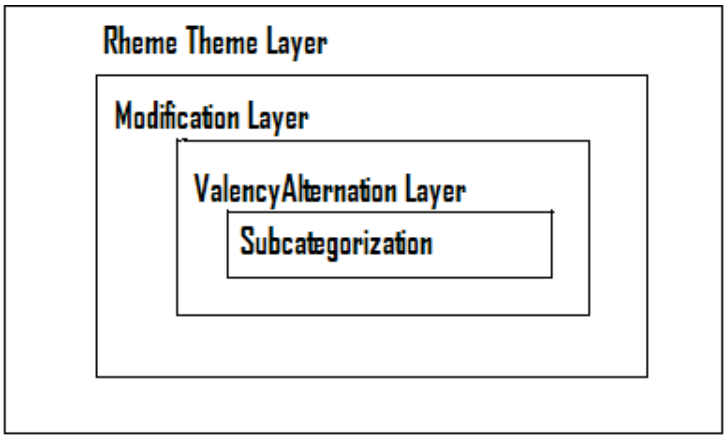

Figure 4 Sketch of Construction Layers

5.2 Constructions in TC

> "Let me start with the following practical example: Suppose I am interested in the following two linguistic topics: serial verb constructions and nominal classification systems: I want to get as much information as possible from those languages that are documented in an on-line archive. If such an archive would be ideal, I could do the following search and get the following kind of information and data. I visit the website… and find a SEARCH function … I type "serial verb construction" and "nominal classification". The search machine presents me the results of the search listing the languages and the files." (Senft 2002, page 3)

Since every token is annotated for language and construction, a search as envisioned in the quote above is fully possible already in TC$_{BETA}$ (see Figure 5). At present the system hosts glossed sentences illustrating, for example, Locative Inversion in Runyankore Rukiga, a Bantu language, and in Lule Sami; nominal constructions in S□kp□le, a Kwa language spoken in Ghana, and periphrastic aspect in Norwegian. All examples can be identified by reference number. They are published together with possible comments from the annotator.

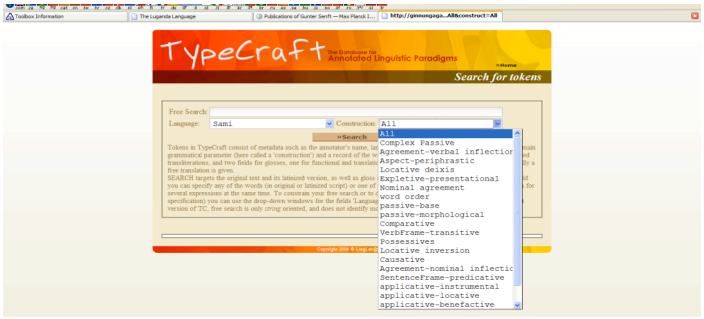

Figure 5: SEARCH for Language and Construction

A different, more difficult, task is how to indicate sets of examples, for example all sentences illustrating Complex Passives in Norwegian. Annotators have found different solutions, for example, some annotators have provided tokens belonging to the same constructional paradigm with partly the same comment in the comment field. The effect is that for each individual token, the main constructional properties are repeated as an additional comment. Here is an example of this pattern:

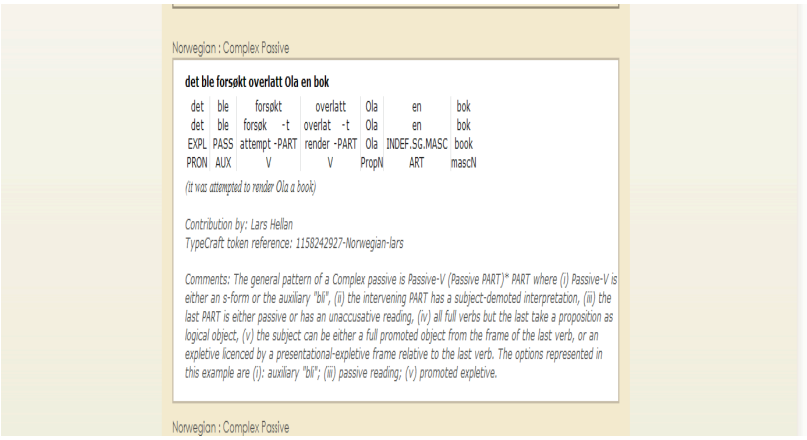

Figure 6 Complex Passives in Norwegian – a token sentence identified in the comment as a member of a constructional type.

A further attempt to preserve construction type information, in those cases where sentences exemplify different subtypes of a construction type, is the naming of the construction (done in the construction field). In Figure 7 several subtypes of applicative constructions are identified via hyphenated names.

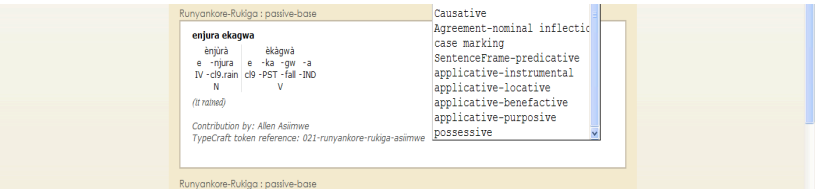

Figure 7 Subtypes of constructions indicated by construction name

Finally a third way to indicate constructional dependency is to introduce a cross reference marker in the comment field of one example, thus indicating the dependency to another token in the base. Following our general policy not to prescribe standards, also when it comes to construction labels, we do not enforce a system of construction types, for example, along the lines suggested here. TC$_{GAMMA}$ instead illustrates a layered network of construction types without hardwiring them in the system.

6 Conclusion

Language preservation and language presentation are two of the central concerns of modern linguistics. With TC$_{BETA}$ we present an online tool for glossing and a public access point to linguistically annotated natural language data. TC$_{BETA}$ seeks to establish a community site for the glossing of sentences with routines that not only provide easy access to the relevant glossing conventions, but also assure that each annotated token is tagged with the relevant meta-information necessary for its retrieval and archiving. The basic data type in TC are annotated sentential and phrasal strings, representing construction types. Also in TC$_{GAMMA}$, which at present is under de-

velopment, the representation of construction information will remain a salient property, the other line of development concern the further development of TC internal communication. Only through extensive collaboration will the linguistic community be able to meet the future demands for richly annotated natural language data.

References

Dorothee Beermann, Lars Hellan and Jonathan Brindle. 2006. TypeCraft: a natural language database; paper presented at the Legon-Trondheim Linguistics Project Meeting in Accra.

Bender, Emily M., Dan Flickinger, Jeff Good and Ivan A. Sag. 2004. Montage: Leveraging Advances in Grammar Engineering, Linguistic Ontologies, and Markup for the Documentation of Underdescribed Languages. *Proceedings of the Workshop on First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation, LREC 2004*, Lisbon, Portugal.

Berge, Anna. 2004. Lectures on Endangered Languages 2: From Kyoto Conference 2002, Review in: International Journal of American Linguistics (69) o4.

Bird Steven & Gary Simons. 2003, Seven Dimensions of Portability for Langugage Documentation and Description. *Language 79*.

Drude, Sebastian. 2002. Advanced Glossing - a language documentation format and its implementation with Shoebox. Talk at the LREC-Workshop in May 2002, Las Palmas.

Good, Jeff . 2006. The ecology of documentary and descriptive linguistic work. Proceedings of the EMELD Workshop 2006: Digital Language Documentation: Tools and Standards: The State of the Art. Lansing, Michigan. July 20–22, 2006.

Lehmann, Christian. 2004. Interlinear morphemic glosses. Manuscript, University of Erfurt, Germany.

Lieb, Hans-Heinrich & Sebastian Drude 2000. Advanced Glossing: A Language Documnetation Format DOBES internal Working Paper.

Senft 2002c "What should the ideal online-archive documenting linguistic data of various (endangered) languages and cultures offer to interested parties? Some ideas of a technically naive linguistic field researcher and potential user", in: Peter Austin, Helen Dry, and Peter Wittenburg, eds. Proceedings of the International LREC Workshop on Resources and Tools in Field Linguistics, Las Palmas, 26-27 May 2002. 15-1-15-11. Las Palmas: European Language Resources Association. <http://www.mpi.nl/lrec>.

Links

'Ailla' databank (http://www.ailla.utexas.org/site/welcome.html)

EMELD project: (http://emeld.org/index.cfm)

Leipzig Glossing Convention: http://www.eva.mpg.de/lingua/files/morpheme.html

Odin language database: (http://www.csufresno.edu/odin)

Summer Institute of Linguistics: http://www.sil.org/

TEI- Text Encoding Initiative: http://etext.virginia.edu/standards/tei/teip4

Toolbox: http://www.sil.org/computing/catalog/