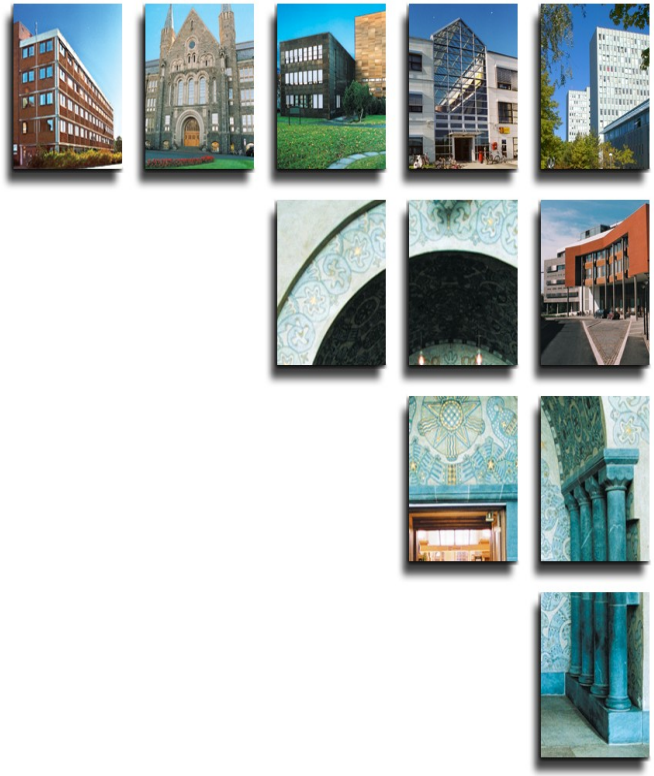


NTNU

Innovation and Creativity

Digital methods for Linguists

Dorothee Beermann

NTNU, Trondheim, Norway

Why do linguists want to work with digital tools

In primary research (Fieldwork) an e-tool can help with the management of linguistic material.

For publication an e-tool can help to create re-usable interlinear glossed examples (IGTs).

In 'empirical phonology' an e-tool can help with signal annotation.

In lexicography an e-tool can help with the creation of dictionaries.

In anthropological studies of language an e-tool can help with the management of audio-based material.

What does that all have to do with Language Description & Documentation

and what is that anyway?

To do LDD means to a comprehensive study of mostly endangered or less-resourced languages. Modern LDD cannot be done without digital tools since it entails the handling of different data types. It further requires that you commit the material that you create to a public archive.

The digital management of electronic language data, however is something that every linguist independent of his/her affiliation does.

Sharing of linguistic data

A new concept which is compatible with LDD but also with all other linguistic approaches and frameworks

We all use mobiles we might even be on Facebook or part of a net-based professional network.
We use e-mail to communicate like never before

But when it comes to the real-time sharing of research data most of us never really thought about it.

However some people did! It is called e-research and its goal for linguistics is to allow a better access to structured language data Archiving is one important thing – active sharing of research data another. We need both!

Overview

* Introduction

*What is Language Description and Documentation ?

*Linguistic methods

*Real-time data sharing

*Uses of real-time data sharing

* * linguistic language promotion

* * linguistic language teaching

*Conclusion

Language Description and Documentation (LDD) is a **new paradigm in linguistics**

~1990 **a computational trend:**
Building language resources is too expensive,
data must be re-usable
Computational resources also for linguists (Bird, Gibbon)

~1998 **a linguistic trend** within functional and descriptive linguistics:

Himmelmann 1998, Evans & Sasse 2003, ...

Trends and Questions

A trend in linguistics:

language endangerment

documenting a language

What is 'data' in linguistics?

gathering data

archiving data

presentation of data

...could that also be something for theoretical linguists?

What is language "endangerment"?

Languages come and go → **language change**

This is normal. However, now change is rapid, and due to globalization, cultures are overrun and languages die.

To be endangered as a language means:

- the speaker community is small
- the language is no longer used to express everything; it becomes "degraded"
- the young generation no longer wants to speak the language.

Language Documentation

"Comprehensive presentation of a language"

A: documentation of a culture (Lehmann, 2001)

Linguistic anthropology with focus on primary data collection (Himmelmann 1998)

B: Comprehensive Language description is not necessarily the same as (only) focusing on primary data collection

Focus on Linguistics

Comprehensive Language description

descriptive + formal as well as quantitative methods

Formal methods:
* models
* notational systems
* computational implementations

Theories as systematic description +
mathematically traceable formalisation

Quantitative methods: data trends and probabilities

Focus on Anthropology

"In early discussions of **language documentation**, the recording of language is generally the primary goal, with work with communities taking a secondary role.
There has been increasing emphasis on community more recently,
with language and linguistics continuing to be at the center in discussions of this extended view of documentation.
Communities are often interested in language conservation, with revitalization frequently part of a broader goal of community development, sustainability, and growth. Where the linguistic notion of documentation fits the community goals is not always clear. "

Rice, K.D. *Strategies for moving ahead: Linguistic and community goals* 2011 - 2nd International Conference on Language Documentation & Conservation.

Why should the normal linguist bother about digital data management?

Inefficiency

Private primary data is often fragmented: bits and pieces of glossed text, partial grammars, some constructions -all somewhere on a PC.

Lack of standards

Uncoordinated transcription conventions, use of proprietary fonts, make-due glosses

Results not falsifiable

Little, scattered data - no means to check the quality of the data

What can be done ?

Linguistic modeling

Language modeling, standardization of grammatical concepts and features can lead to unified standards and an improved uniformity of linguistic resources.

Suitable linguistic tools for language processing

Linguistic tools can lower the technical threshold, so that ordinary working linguists can use modern technology to create and structure linguistic data

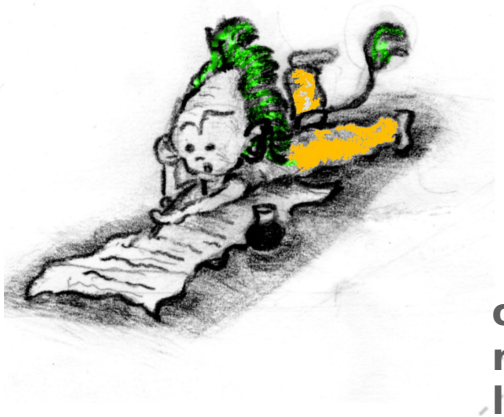
Sharing of linguistic data in collaborative databases

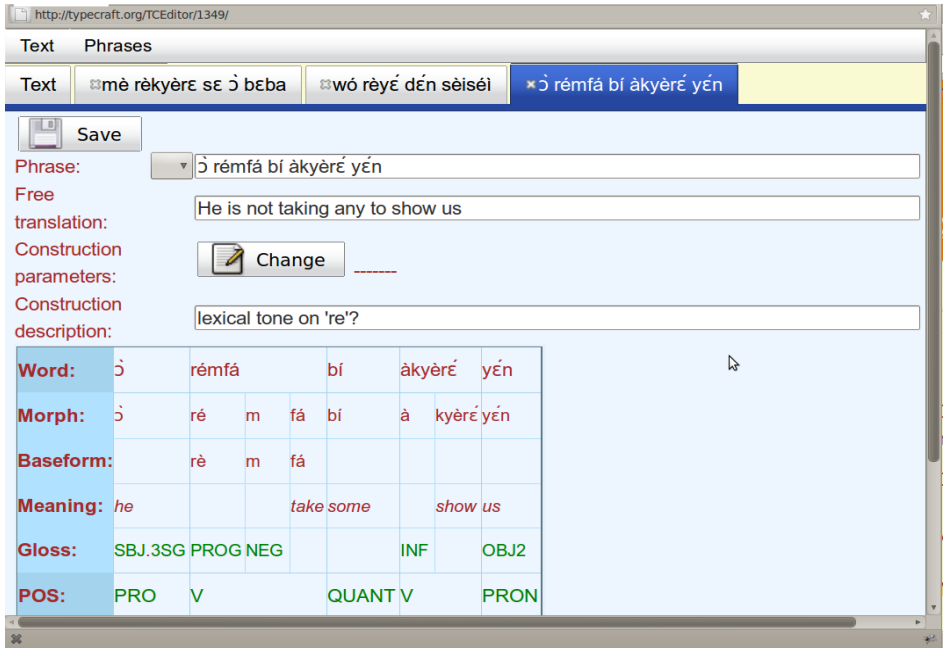
Sharing of information is done online (news, personal information (pictures, opinion) can be found directly online. Why not do the same with research data?

Interlinear Glossed Text

Create, store, retrieve, share

* Interlinear Glosser
* Repository of Interlinear Glossed Text (IGT)
* Collaborative Editing





For
Language Studies in the Humanities
Language Science and Teaching

Linguists
Language Teachers
Anthropologists

online service

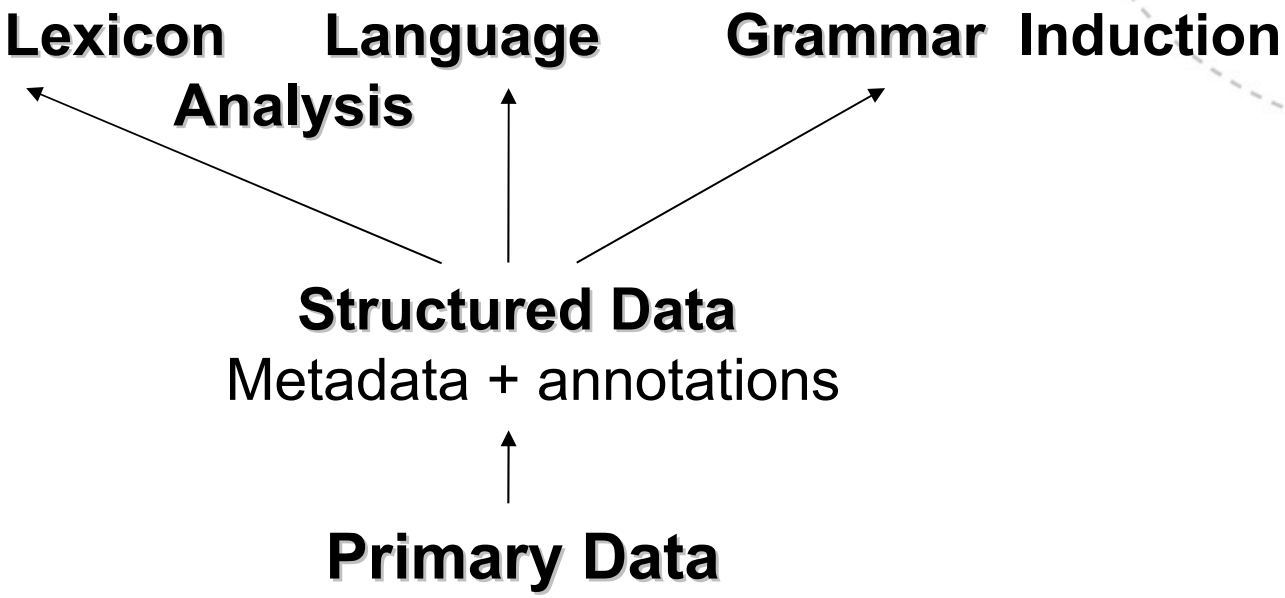
NTNU
Innovation and Creativity

What is data to the ordinary working linguist?

A non-computationally inclined linguist does rarely get the chance to create a multi-million word corpus, instead for most linguists who are data-oriented to be able to build and maintain a 'working corpus' for on-going research and teaching is sufficient. (Austin 2006)

How can a linguist harmonize her/his research goals and the structure of his/her data with his research goal? Which is the most appropriate method to obtain data?

From data to language modeling



audio & textual data

Linguistic procedures

Transcription

Signal annotation

Creation of suitable text material

Text annotation - creation of annotation profiles

Questionnaires and Elicitation (linguistic experiments ?)

Lexicon extraction

Concordancing and signal access via annotation, combination of signal and text annotation

Lexical property extraction

Empirical methods in linguistics

* Introspection
* Experimental methods, interviews, questionnaires
* Corpus methods

Not so clear to which extend each of these methods are quantitative and to what extent the observations we make are authentic

(How can we observe language that speakers produce when they are NOT observed?)

An awareness of linguistic methodology makes us understand the importance of high quality of linguistic data.

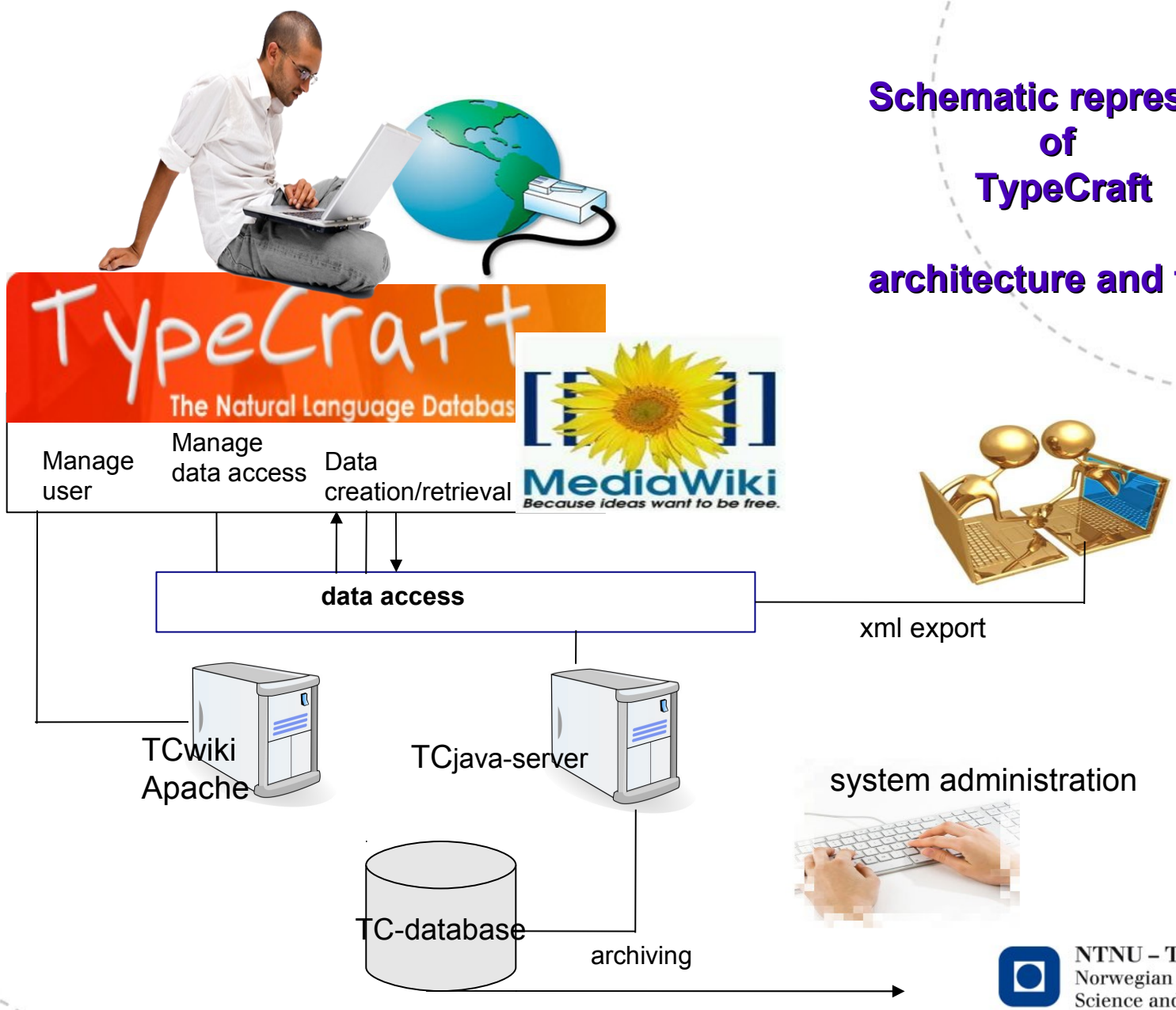
Giving linguistic annotations to our primary data is the first step to linguistic analysis.

Knowing how time-consuming data analysis is and being familiar with the high-level of expertise that is needed to do annotation we start to appreciated how important the sharing of structured linguist data is.

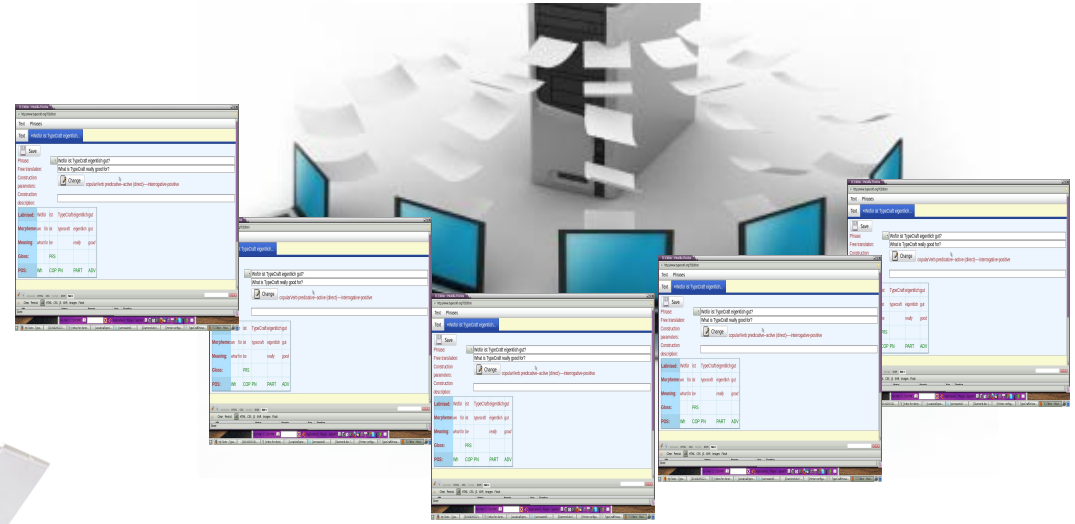
Let's look at one way how data sharing can be done.

Schematic representation of TypeCraft

architecture and functions



Interlinear Glossed Text Brokerage



NTNU
Innovation and Creativity

TypeCraft Storage and Datamodel

```xml
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE phrases PUBLIC "-//LINGLAB//DTD TC PHRASES 1.0//EN" "http://www.typecraft.org/dtd/tc1.dtd">
<?xml-stylesheet type="text/xsl" href="tcphrases.xsl"?>

<phrases>
  <phrase id="18659" valid="VALID">
    <original>ò àkyérέw ǹhómá nò</original>
    <translation>He has written the letter</translation>
    <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="ò">
      <pos>PRO</pos>
      <morpheme id="111636" text="ò" baseform="">
        <gloss>3SG</gloss>
      </morpheme>
    </word>
    <word id="68410" text="àkyérέw" head="yes">
      <pos>V</pos>
      <morpheme id="111637" text="à" baseform="à">
        <gloss>PERF</gloss>
      </morpheme>
      <morpheme id="111638" text="kyérέw" baseform="kyérέw" 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>PRO</pos>
      <morpheme id="111640" text="nò" baseform="nò">
        <gloss>3SG</gloss>
      </morpheme>
    </word>
  </phrase>
```

Fig. 5 XML export from TypeCraft

TC uses an PostgreSQL database for data storage.

The data mapping between Java objects and database tables is managed by Hibernate. TC is not bound to any specific SQL database.

TypeCraft data can be divided into two specific types:

Common data: pos tags, gloss tags, global tags, ISO 639-3 languages
Shared between all annotated tokens and users.

Individual data: texts, phrases, words and morphemes, together with their annotation. This is data specific to each user.

Individual data items reference common data items.

How can data sharing be used ?

One important user group **African Linguists**

NO CORPORA

→ create language resources

LITTLE BOOKS AVAILABLE

→ make them accessible to others

"Add my voice by describing my language"

University of Ghana, Legon

EDUCATIONAL POLICY

→ draw attention to my language

NO PUBLICATION CHANNELS

→ make my work available

NTNU – Trondheim
Norwegian University of Science and Technology

Two years for a master in Linguistics !

TypeCraft

The Natural Language Database

classroom | discussion | view source | history

Classroom:The word KU in Runyankore-Rukiga

This page was created as an in-classroom exercise in LING 2208, NTNU

Author: Franciane Rocha (Last one in the picture)

Author: Misah Natumanya (First one in the picture)

This page is about the analysis of the grammatical function of the word "ku" in Runyankore-Rukiga.

The data is the result of Typecraft's phrase search done on March 18, 2011.

Contents [show]

Interlinear Glossed Text

the root of all linguistic research

group of LING 2208 (Linguistics) spring 2011 missing Prof. Dorothee Beermann

The conjunction KU in Runyankore-Rukiga

Generalization: "KU as a word in Runyankore-Rukiga works as a conjunction".

We found 46 phrases in which Ku had the function of a conjunction. Some of them were glossed only as CONJ and others had more specific information and were glossed as subordinative conjunctions or as one of its subcategories namely complementizer, relativizer and adverbializer.

In this research we used the definitions about conjunction and its subcategories provided by the Glossary of Linguistic Terms of SIL - Summer Institute of Linguistics [1]. The relationship between conjunctions and its subcategories found in SIL's glossary can be summarized in the table below, where **the bolded elements** and their disposition across the table, show the elements and the relationships we observed in our data set. It is important to point out that no examples were found of the word KU working as a coordinative conjunction or its subcategory correlativizer. This observation allowed us to draw a more specific generalisation, namely, that the word KU works as a **subordinative** conjunction in Runyankore-Rukiga.

Conjunction			
SUBORDINATIVE			Coordinative
COMP	**REL**	**ADVL**	Correlativizer

"Recently linguistic data has come under scrutiny. Researchers from different linguistic fields have questioned its validity, and the integrity of theories that "are built" on this data."

Ku as only CONJ

30 of our tokens are glossed only as conjunction (CONJ) As in:

Ku naahikire ahari Butunduuzi, emotoka ereemerera obwe ku baabarareebaga emotoka ereemerera baija kureeba omufu!

"When i arrived at Butunduuzi, and then whenever the car stopped, they would come to see the dead-person."

Ku	naahikire	ahari	Butunduuzi	emotoka	ereemerera	obwe	ku	baabarareebaga	emotoka	ereemerera	baija	kureeba	omufu!
ku	n aa hik ire	a hari	butunduuzi	e motoka	e ra emerer a	obwe	ku	ba a ba ra reeb a ga	e motoka	e re emerer a	ba ij a	ku reeba	o mu fu!
when	1SG PASTRE *reach* PERF	IV *at.*SPTL		IV *car*	AGR ASP *stop* FV	*then*	*when*	3PL PASTRE 3PL ASP *see* FV EMPH	IV *car*	AGR ASP *stop* FV	3PL *come* FV	INF *see*	IV CL1 *the-dead*
CONJ	V	PREP/PROSPT	PN	N	V	CONJ	CONJ	V	N	V	V	V	

KU as a CONJS

Which linguistic tools are available?

	Tools	Archiving	Presentation	Processing
Text	Toolbox,	external databases	Ø	Ø
	Flex,		PDF, html	XML
	TypeCraft	TypeCraft	in all editors	XML
Audio	Praat,	external databases		
	Transcriber	?	?	
Video	Elan	?	?	

TypeCraft	FLEx
Typologically oriented, easy to switch between languages, integrated ISO-language-lists and transliteration functionality.	Designed for the work with one language
Server solution, web-based browser application, distributive use, propagation of the concept of Open Scientifc Data and a collaborative approach to research.	single-user desktop system
Basic morphological parsing	Morphological parser well-integrated with the tool's lexicon functionality
Main functions: Sharing of linguistic data online (in human-readable and in XML format). Export of interlinear glosses for use in paper publications. Collaborative approach to the creation of re-usable linguistic data and the standardization of linguistic annotations.	Integration of interlinear glossing and lexicon work, export of digital dictionaries.

Table 1 A comparison of some of the key-features of TypeCraft and FLEx

There are different ways of data sharing!

Sharing can be done by:

Archiving in one of the specialised institutional centers, such as
Some funders might require researchers to deposit
their data in an archive managed by the funding institution. Advantages of centralized data centers are better control over standards, data sharing policy and perhaps a better data quality.

Alternative: Self -archiving as part of a shared research infrastructure

+ openness, transparency, flexibility, real-time data sharing

= safe-keeping, long-term preservation, data accessibility

- danger of reduced data quality

References

Austin, P. K. (2006). Data and Language Documentation. In Gippert, J., Himmelmann, N. P., and Mosel, U., editors, Essentials of Language Documentation, chapter 4, pages 87 – 112. Mouton de Gruyter, Berlin/New York.

Evans, Nick and Hans-Jürgen Sasse. 2003 “Searching for meaning in the Library of Babel: field semantics and problems of digital archiving”. In Researchers, Communities, Institutions, Sound Recordings, (eds.) Linda Barwick, Allan Marett, Jane Simpson and Amanda Harris. Sydney: University of Sydney.

Gibbon, D. (2002b). Workable Efficient Language Documentation: a Report and a Vision. ELSNews, 11.3.

Himmelmann, N. (1998). Documentary and Descriptive Linguistics. Linguistics, 36:161–195.

Lehmann, C. (1999). Documentation of Endangered Languages: A Priority Task for Linguistics. Arbeitspapiere des Seminars f˙ur Sprachwissenschaft der Universität Erfurt.

Lehmann, C. (2001). Language Documentation A Program. In Bislang, W., editor, Aspects of Typology and Universals. Akademie Verlag, Berlin.

Lehmann, C. (2002). Structure of a Comprehensive Presentation of a Language. In Tsunoda, T., editor, Basic materials in minority languages 2002, pages 5–33. ELRP Publication Series B003, Osaka: Osaka Gakuin University