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. “

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
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

Lexicon  Language  Grammar  Induction
Analysis

Structured Data
Metadata + annotations

Primary Data
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:

- Manage user
- Manage data access
- Data creation/retrieval
- Data access
- TCwiki
- Apache
- TCjava-server
- TC-database
- Archiving
- XML export
- System administration

NTNU – Trondheim Norwegian University of Science and Technology
Interlinear Glossed Text Brokerage
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, morphemes, together with their annotation. This is data specific to each user.

Individual data items reference common data items.

---

*Fig. 5* XML export from TypeCraft
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”

**EDUCATIONAL POLICY** → draw attention to my language

**NO PUBLICATION CHANNELS** → make my work available
Two years for a master in Linguistics!

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.
### Which linguistic tools are available?

<table>
<thead>
<tr>
<th>Tools</th>
<th>Archiving</th>
<th>Presentation</th>
<th>Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Text</strong></td>
<td>Toolbox, external databases</td>
<td>Ø</td>
<td>Ø</td>
</tr>
<tr>
<td></td>
<td>Flex, TypeCraft</td>
<td>PDF, html</td>
<td>XML</td>
</tr>
<tr>
<td></td>
<td></td>
<td>in all editors</td>
<td></td>
</tr>
<tr>
<td><strong>Audio</strong></td>
<td>Praat, external databases</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td>Transcriber</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td><strong>Video</strong></td>
<td>Elan</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>TypeCraft</td>
<td>FLEEx</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------------------------------------------------</td>
<td>---------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Typologically oriented, easy to switch between languages, integrated ISO-language-lists and transliteration functionality.</td>
<td>Designed for the work with one language single-user desktop system</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Server solution, web-based browser application, distributive use, propagation of the concept of Open Scientific Data and a collaborative approach to research.</td>
<td>Basic morphological parsing Morphological parser well-integrated with the tool’s lexicon functionality</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic morphological parsing</td>
<td>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.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>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.</td>
<td>Integration of interlinear glossing and lexicon work, export of digital dictionaries.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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


