The “Digital Grammar” Project

Integrating the Wiki/CMS approach

with Language Archiving Technology and TEI

SEBASTIAN DRUDE

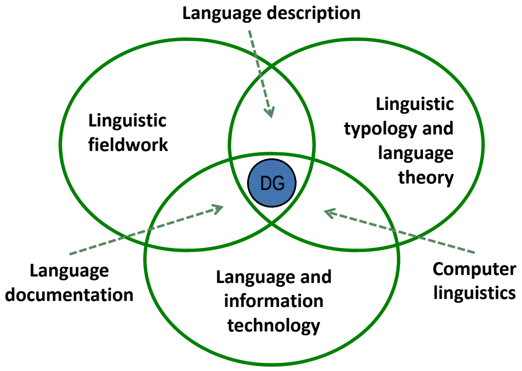
Goethe-Universität Frankfurt (Germany)
& The Language Archive / MPI-PL Nijmegen

TEI-Conference, SIG Linguistics
October 14th, 2011

1. Background

- Project (searching for funding and people)
- Context: “A discourse-based multimedia grammar of Awetí” (Dilthey fellowship)
- Wider context: DOBES-programme
- “Digital Grammars” should be an authoring system, useful to descriptive linguists, typologists etc.
- Community-based project
- Building on previous relevant work
  (some being presented in this colloquium)

1. Background

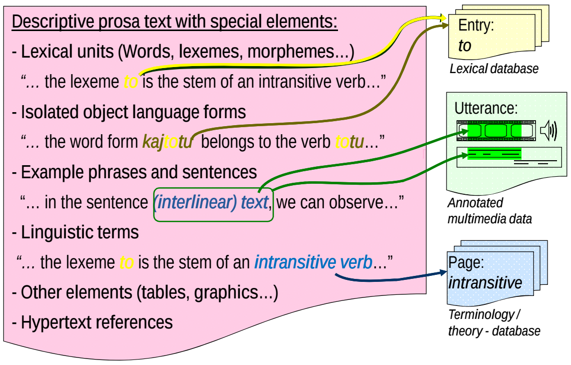


2. “Digital Grammars”

Key features:

- Descriptive texts as digital documents
- Different from office-software or PDFs
- True hypertext documents
  with individual but interlinked pages
- Logical /content mark-up with functionalities instead of visual formatting
- Questions: - Authoring/editing software?
  - Archiving, working, display data formats?
  - Different uses (humans / machine)?

2. “Digital Grammars”



Descriptive prosa text with special elements:

- Lexical units (Words, lexemes, morphemes…)
  “… the lexeme *to* is the stem of an intransitive verb…”
- Isolated object language forms
  “… the word form *kajtotu* belongs to the verb *totu*…”
- Example phrases and sentences
  “… in the sentence *(interlinear) text*, we can observe…”
- Linguistic terms
  “… the lexeme *to* is the stem of an *intransitive verb*…”
- Other elements (tables, graphics…)
- Hypertext references

Entry:
to
Lexical database
Utterance:
Annotated multimedia data
Page:
intransitive
Terminology / theory - database

2. “Digital Grammars”

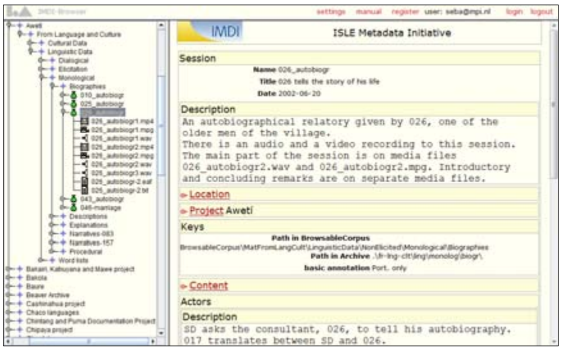
Principal functionalities:

- Links from *exemplars* to multi-media utterances in a corpus
- Automatic generation of concordances
  for more *examples*
- Lexical units are linked to an online-lexicon
- Separation of description from theoretical parts (e.g., explanation of analytical concepts and theory in a terminological database)
- Different versions for different audiences instead of footnotes

3. Language Archiving Technology

- Developed at the Max-Planck-Institute for Psycholinguistics since ca. 2000
- Context: language documentation (DOBES)
- Language corpora with IMDI-metadata
- ELAN and ANNEX for creating / viewing annotated multimedia-data
- LEXUS for online lexical databases
- ISOcat data catalogue for linguistic terms
- No component for scientific meta-texts such as typological work or grammars

3. Language Archiving Technology:
IMDI - MetaData tree (online-archive)



ISLE Metadata Initiative

Session

Name 026_autobiogr
Title 026 tells the story of his life
Date 2002-06-20

Description

An autobiographical relatory given by 026, one of the older men of the village. There is an audio and a video recording to this session. The main part of the session is on media files 026_autobiogr2.wav and 026_autobiogr2.mpg. Introductory and concluding remarks are on separate media files.

Location

Project Aweti

Keys

Path in BrowsableCorpus BrowsableCorpus\MatFromLangCult\LinguisticData\NonElicited\Monological\Biographies
Path in Archive .\fr-lng-clt\lng\monolog\biogr\
basic annotation Port. only

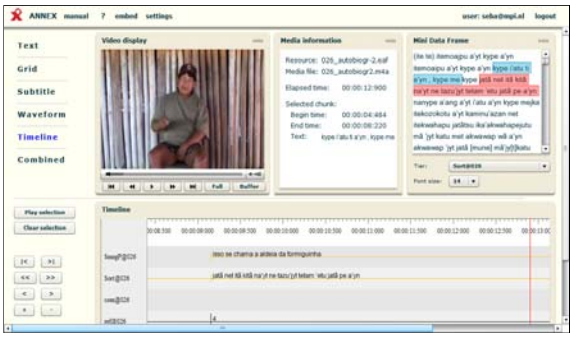
Content

Actors

Description

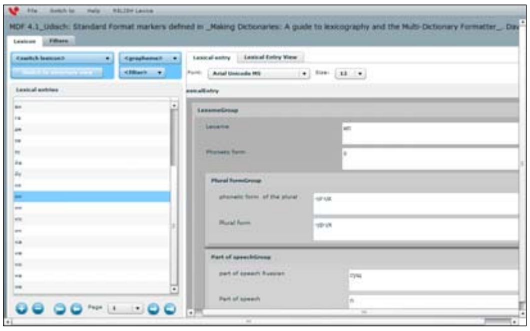
SD asks the consultant, 026, to tell his autobiography. 017 translates between SD and 026.

3. Language Archiving Technology:
Annotated Text in Annex (Elan Ann. Format)



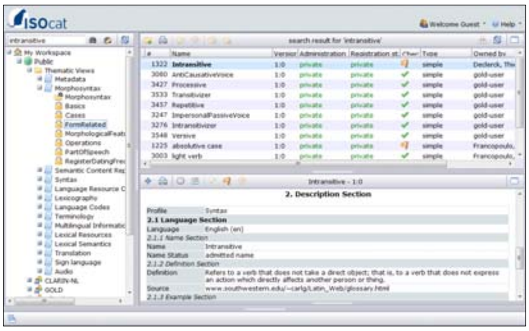
Resource: 026_autobiogr-2.eaf
Media file: 026_autobiogr2.m4a
Elapsed time: 00:00:12.900
Selected chunk:
Begin time: 00:00:04.484
End time: 00:00:06.220

3. Language Archiving Technology:
LEXUS lexical online database



MDF 4.1_Udach: Standard Format markers defined in _Making Dictionaries: A guide to lexicography and the Multi-Dictionary Formatter_

3. Language Archiving Technology:
ISOcat data category (terminology) registry



search result for 'intransitive'

#	Name	Version	Administration	Registration st.	Chec	Type	Owned by
1322	**Intransitive**	1:0	private	private		simple	Declerck, The
3060	AntiCausativeVoice	1:0	private	private	✓	simple	gold-user
3427	Processive	1:0	private	private	✓	simple	gold-user
3533	Transitivizer	1:0	private	private	✓	simple	gold-user
3457	Repetitive	1:0	private	private	✓	simple	gold-user
3247	ImpersonalPassiveVoice	1:0	private	private	✓	simple	gold-user
3276	Intransitivizer	1:0	private	private	✓	simple	gold-user
3548	Versive	1:0	private	private	✓	simple	gold-user
1225	absolutive case	1:0	private	private		simple	Francopoulo,
3003	light verb	1:0	private	private	✓	simple	Francopoulo,

Intransitive - 1:0

2. Description Section

Profile	Syntax
2.1 Language Section	
Language	English (en)
2.1.1 Name Section	
Name	Intransitive
Name Status	admitted name
2.1.2 Definition Section	
Definition	Refers to a verb that does not take a direct object; that is, to a verb that does not express an action which directly affects another person or thing.
Source	www.southwestern.edu/~carlg/Latin_Web/glossary.html
2.1.3 Example Section	

4. The Wiki/CMS approach

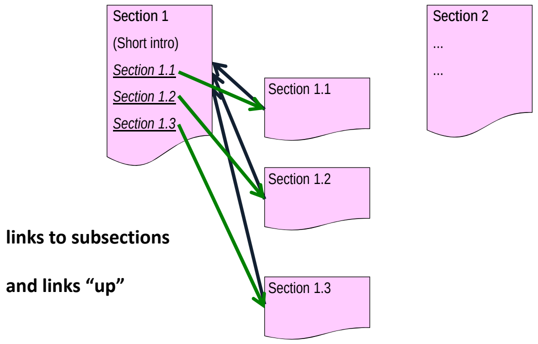
Content Management Systems & Wikis:

- Online-collaboration
- User-management
- Version control
- Updates etc. are developed by others

Challenges:

- Serial & hierarchical ordering of pages
- Specialized markup (extensible)
- Special functionalities

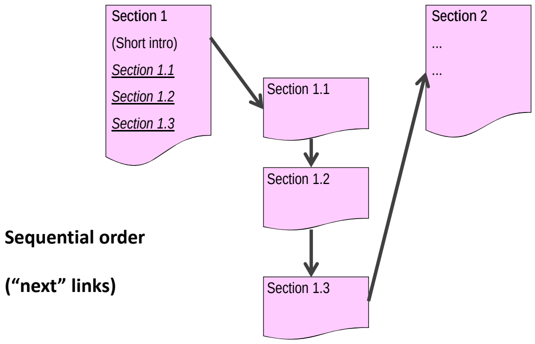
4. The Wiki/CMS approach



links to subsections

and links "up"

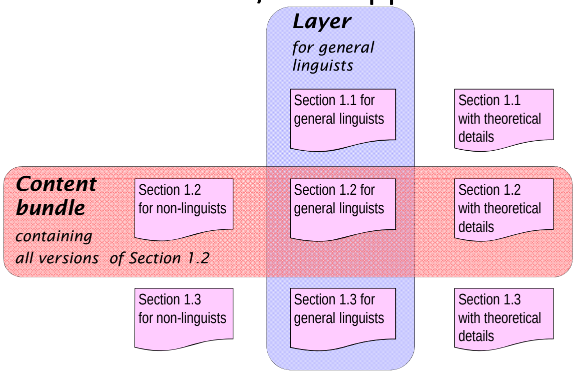
4. The Wiki/CMS approach



Sequential order

("next" links)

4. The Wiki/CMS approach


Layer
for general linguists
Section 1.1 for general linguists
Section 1.1 with theoretical details
Content bundle
containing all versions of Section 1.2
Section 1.2 for non-linguists
Section 1.2 for general linguists
Section 1.2 with theoretical details
Section 1.3 for non-linguists
Section 1.3 for general linguists
Section 1.3 with theoretical details

5. Text Encoding Initiative (XML)

- XML promises to be a lasting standard
- Human and machine readable
- The TEI "recommendations" build on XML
- TEI is a widely used *de-facto* standard
- TEI-XML should be one format (archival, interchange) of the digital grammar

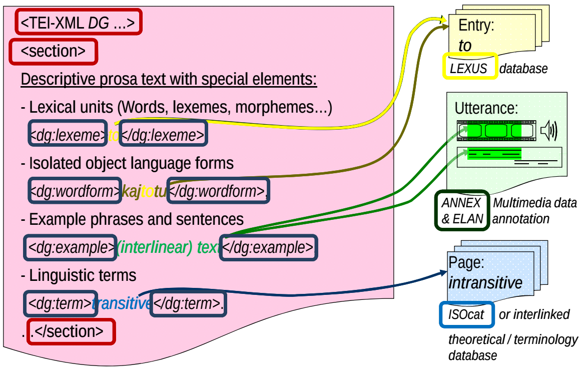
Challenges:

- How to enter (XML) text & markup?
- No specific TEI module for linguistic descriptive / typological work yet

5. Text Encoding Initiative (XML)

Linguistic / ontological Type	Tags and properties	Formatting	Main functionality	Possible secondary functionalities (tooltips and the like)
syntactic unit (sequence of words)	<dg:SUnit > he goes </dg:SUnit >	***he goes*** italics, roman	play media file	• see interlinear glosses • see syntactic tree • links to lexical entries for individual words
single word	<dg:word> goes </dg:word>	***goes*** italics, roman	link to lexical entry	• see interlinear glosses • play media file if exists
lexical word	<dg:lexwd homn.nb =1 > go </dg:lexwd >	***go***$_1^{W}$ italics, roman, superscript W,	link to lexical entry (select if there are homonyms)	• show meaning • show word class

5. Text Encoding Initiative (XML)



<TEI-XML *DG* ...>

<section>

Descriptive prosa text with special elements:

- Lexical units (Words, lexemes, morphemes...)

<dg:lexeme>to</dg:lexeme>

- Isolated object language forms

<dg:wordform>kajtotu</dg:wordform>.

- Example phrases and sentences

<dg:example>(interlinear) text</dg:example>

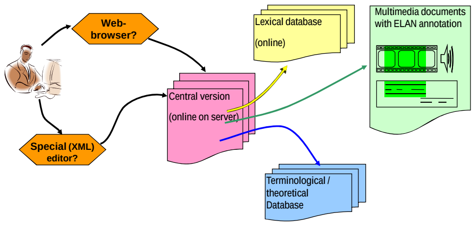
- Linguistic terms

<dg:term>transitive</dg:term>.

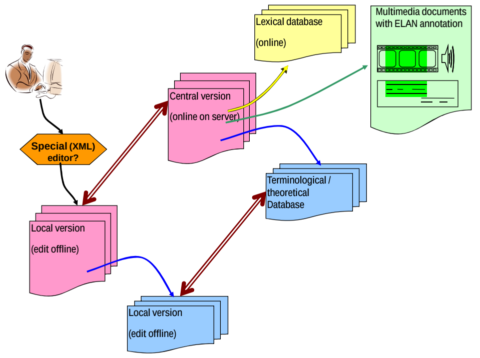
..</section>

Entry:
to
LEXUS *database*
Utterance:
ANNEX & ELAN *Multimedia data annotation*
Page:
intransitive
ISOcat *or interlinked*
theoretical / terminology database

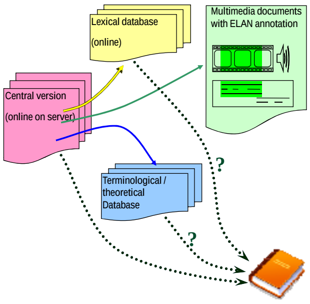
6. Versioning and publication



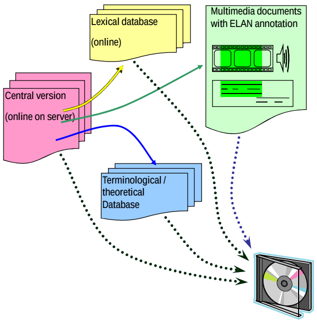
6. Versioning and publication



6. Versioning and publication


Central version
(online on server)
Lexical database
(online)
Multimedia documents with ELAN annotation
Terminological / theoretical Database
?
?

6. Versioning and publication


Central version
(online on server)
Lexical database
(online)
Multimedia documents with ELAN annotation
Terminological / theoretical Database

6. Versioning and publication

- DGs are not static but “living documents”
- Version control (automatic in a Wiki/CMS)
- Printing the grammar should be possible
- and producing digital “snapshot” distributions
- Versions and distributions must be citable
- Ideally, it would be possible to work on an offline copy (e.g., in the field)
- This posits complex issues of synchronization
- and, again, the question of a suitable editor

7. Conclusion

- The project needs first to decide on some basic questions
- … to learn about existing standards (TEI) & tools
- … and to establish realistic goals and priorities
- This is work which needs input from and exchange with a community of experts
- Group and individual meetings planned
- I hope that this colloquium helps to form such a community
- Comments are welcome! Thank you!