LEXUS and ViCoS

Jacqueline Ringersma

The Language Archive
Max Planck Institute for Psycholinguistics

DGfS-CNRS Summer School on Linguistic Typology
LEXUS & ViCoS: From lexical to conceptual spaces

**LEXUS**: a web based tool for the creation of multimedia encyclopedic dictionaries and lexica

**ViCoS**: a web based tool for constructing and visualizing conceptual spaces. Enrichment of lexical information with conceptual knowledge
Structure
Structure of lexicon depends on language, linguistic theory, purpose

Conceptual
Large variation in linguistic concepts (attribute) and value naming

Format
Large variation in formats (XML, Shoebox, Chat, Word)

Data interoperability problem
cross lexica searches, merging, comparison

Archive requirements
Representation format (XML)
One archive exploitation framework
LEXUS development
2005: First version of LEXUS developed with some input from DoBeS
2006: Construction of Yeli Dnye multimedia lexicon (Steve Levinson)
2007: Start of Marquesan and Tuamotuan dictionary project
   (Gaby Cablitz, Edgar Tetahiotupa, Jean Kape, others)

- Many improvements on the LEXUS functionalities
- Many improvements on interoperability with Toolbox and XML
- Initial development of ViCoS, based on ideas from Gaby Cablitz and the Tuamotuan participants in the project
- Input on the new LEXUS user interface from the Marquesan and Tuamotuan participants in the project
LEXUS & ViCoS: Development

2007: Start of Marquesan and Tuamotuan dictionary project (Gaby Cablitz, Edgar Tetahiotupa, Jean Kape, others)

- Many improvements on the LEXUS functionalities
- Many improvements on interoperability with Toolbox and XML
- Initial development of ViCoS, based on ideas from Gaby Cablitz and the Tuamotuan participants in the project
- Input on the new LEXUS user interface from the Marquesan and Tuamotuan participants in the project
2007-2008: Input from other DoBeS projects to LEXUS
2008-2009: Further refinements of ViCoS
New LEXUS user interface
2009: Launch of the tools, Workshop April 2, 2009

2006→2009: Do we (&you) still need these tools?

Christopher Doty: Bridging the Gap between Linguistics and Community: Producing materials for language maintenance

Ulrike Mosel: Turning the linguist's lexical database into a community dictionary “the differences between a lexical data base as it typically results from a language documentation project and the kind of dictionaries the speech community wants”
LEXUS technicalities
LEXUS & ViCoS: Technicalities

Based on two ISO TC 37 standards for linguistic resources

**LMF**: Lexical Markup Framework (lexicon structure)

**DCR**: set of standardized data categories to be used as a reference for the definition of linguistic annotation schemes or any other formats used in the area of language resources (concept naming)

**LMF/DCR:**
- A modular structure for content interoperability between lexical resources.
- Archiving exploitation framework, XML based
Lexical Markup Framework: defines the default structure of a lexicon

The lexicon schema is seen as lexical attributes (data categories) grouped together with others (data category groups) and embedded in a tree structure.

A lexicon needs a structure first. Content can be built on this structure.
**LexicalEntry**: container for managing one or several forms and possibly one or several meanings in order to describe a lexeme

**Lexeme**: abstract unit, generally associated with a set of forms sharing a common meaning

**Form**: text string representing the word

**Sense**: specifies the meaning and context
LEXUS & ViCoS:
Technicalities
Data category registry:

A set of standardized data categories to be used as a reference for the definition of linguistic annotation schemes or any other formats used in the area of language resources (concept naming)

A modular structure for content interoperability between lexical resources → this facilitates searching cross lexica, sharing & merging

Archiving exploitation framework, XML based
LEXUS & ViCoS: Technicalities
LEXUS functionalities
LEXUS Functionalities

• Creation of lexica from scratch, import lexica from other formats (Toolbox, XML, Chat)
• User definable views of word list and lexical entries
LEXUS Functionalities

- Creation of lexica from scratch, import lexica from other formats (Toolbox, XML, Chat)
- User definable views of word list and lexical entries
- Linking multi media fragments to lexical entries
LEXUS Functionalities

- Creation of lexica from scratch, import lexica from other formats (Toolbox, XML, Chat)
- User definable views of word list and lexical entries
- Linking multi media fragments to lexical entries
- Setting sort orders at workspace level
LEXUS Functionalities

- Creation of lexica from scratch, import lexica from other formats (Toolbox, XML, Chat)
- User definable views of word list and lexical entries
- Linking multimedia fragments to lexical entries
- Setting sort orders for each data category
- Using different character sets (UTF-8 is required)
- Input of non-Latin characters
LEXUS Functionalities

- Creation of lexica from scratch, import lexica from other formats (Toolbox, XML, Chat)
- User definable views of word list and lexical entries
- Linking multi media fragments to lexical entries
- Setting sort orders for each data category
- Using different character sets (UTF-8 is required).
- Input of non-Latin characters
- Search all lexica in the workspace or
- Setting filters on the word list of one lexicon
LEXUS Functionalities

- Creation of lexica from scratch, import lexica from other formats (Toolbox, XML)
- User definable views of word list and lexical entries
- Linking multi media fragments to lexical entries
- Setting sort orders for each data category
- Using different character sets (UTF-8 is required)
- Search all lexica in the workspace or
- Setting filters on the word list of one lexicon
- Access rights (read/write) for registered users
- New: import to existing lexica

Demo of LEXUS
ViCoS

Lexus:
- Accessible online
- Rich in multimedia
- Adjustable display

Still it is a list of isolated entries
• Existing lexical resource is starting point
  – Words offer key access

• ViCoS is technology to link words – and the associations they evoke – to other words

• Ease-of-use to empower community members to
  – actively describe their language & culture
  – and to learn from such resources
  → resources targeted for human consumption

• Resulting Conceptual Spaces = sort of informal ontology of fuzzily-defined concepts and relationships
  – but where concept nodes are anchored in corresponding formal resources (lexica)
ViCoS demo
The Iwaidja language of the Cobourg Peninsula, Northern Territory, Australia (approximately 200 speakers)

Hans-Jorgen Sasse - Koln
Nick Evans - Melbourne
Linda Barwick - Sydney
Bruce Birch – Melbourne
Dr Murray Garde - Jabiru
Joy Williams - Minjilang
Janet Fletcher - Melbourne

More information on http://www.mpi.nl/DOBES/projects/iwaidja
Minjiang Endangered Languages Publication (MELP) project

Lexus multi media lexicon

ViCoS conceptual spaces

More involvement of the speech community

The creation of an accessible and attractive online dictionary, with multimedia capability and cross-referenced conceptual spaces will contribute to conservation, revitalization and empowerment of the language and community.
Toolbox lexicon:
3500 entries, still growing
Multiple people contribute
Structure comes from *.typ file

Content comes from *.lex (or *.dic) file

*.typ file consists of marker definitions

\+mkr ps
\nam part of speech
\lng Default
\mkroVerThis lx
\-mkr

\+mkr sn
\nam sense number
\lng Default
\mkroVerThis ps
\-mkr
Structure comes from *.typ file

Content comes from *.lex (or *.dic) file

*.lex file consists of lexical entries

```
\lx ababa
\lc ababa
\ps n
\de stuttering
\ge stuttering
\dt 26/Dec/2006
```

```
\lx abarrk
\lc abarrk
\ps n.
\de cut of dugong meat taken from side of belly
\ge cut_of_dugong_meat
\dt 09/Dec/2005
\sd dugong
\sd butchering
```
Problem!

Toolbox is too liberal in allowing the variations in the structure of the lexical entries.

No need to be consistent with the *.typ file.

For the human eye: no problem.

but: computers cannot make interpretations.
Toolbox2LEXUS

Wara warrkbi abangkun.
There’s a big, tall man.

This contrasts with walarl ‘tall and skinny, gangly’.

body
2
father
father

Kudnayanjing janad rimardyarrwun ba wara, abangkun ba wara, riki bani marr
You see that bloke there walking, that’s his father, this one here is his son.

abNkun
169, 171

father emu
father_emu
bird

Warlkbuyiŋ kayidaruny ba kurlajuk, abangkun rajaraka kurlajuk.
(The mother emu) lays the eggs and leaves them, the father emu looks after the eggs.
<table>
<thead>
<tr>
<th>\lx</th>
<th>\lc</th>
<th>\sn</th>
<th>\ps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>n</td>
</tr>
</tbody>
</table>

Beach hibiscus. Rope for harpoons and tying up canoes is made from this tree species, and the timber is used to make fish names.

**alabanja**

**habiscus, beach**

205,410, IE 84

**plant**

**material**

**Iwa05.Feb2**

**alabanja alhurdu**

**hibiscus string/rope**

**short-finned batfish**

**Zabidius novaemaculatus**

**animal**

**fish**

**Iwaidja Fish Names.xls**

**MELP project elicitation**

SH

19/Dec/2006

Language consultants identify other fish species which have a moon-like shape as alabanja (e.g. juvenile of diamond trevally).
Solution: chunking!
Toolbox2LEXUS

Getting Toolbox into LEXUS

Solution: chunking!
Toolbox2LEXUS

Show structure
Toolbox2LEXUS