Online lexicons, LEXUS and ViCoS

Jacqueline Ringersma
Huib Verweij
Marc Kemps-Snijders
Online lexicons: a general and brief introduction

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
Lexicon:

Works, made for humans, about words
Main content is divided into lexical entries, each of which is about a word

Examples:
Standard dictionaries
Bi-lingual dictionaries (=maybe 2 lexicons?)
Ethnographic dictionaries
Online lexicons: what’s different?

The design of online lexicons
Burke, Sean Michael
1998

Online:

Move from product (book) to a service:
one single database (vs. many copies)
supply services (access, search, customization)
Online dictionary **MACRO STRUCTURE:**

Printed form: designed arrangement entries via sorted word list (alphabet)

Online form: **procedural structure** how the user can access the lexicon

word list (any sorting)
search engine options (=, b, e, c)
filtering
fuzzy matching
Online lexicons: what’s different?

Online dictionary **MACRO STRUCTURE**:

Requires a well thought of lexicon structure!

**dog** \d\o.g\ n. *often attrib* [ME, fr. OE docga] 1\a\: a highly variable carnivorous domesticated mammal (*Canis familiaris*) prob. Descended from the commonn wolf; broadly : any animal of the dog family (*Canidae*) to which this mammel belongs **b**: a male dog 2\a\: a worthless fellow : **b**: CHAP, FELLOW <a gay ~> 3\a\: any of various usu. simple mechanical devices for holding, gripping, or fastening consisting of a spike, rod or bar **3b**: ANDRION 4\a\: SUN DOG 4\b\: WATER DOG 4\c\: FOGBOW **dog-like** \d\o.gli-k\ adj
**dog** \do\n. **often attrib** [ME, fr. OE docga]  

1a: a highly variable carnivorous domesticated mammal (*Canis familiaris*) prob. Descended from the common wolf; broadly : any animal of the dog family (*Canidae*) to which this mammel belongs  

b: a male dog  

2a: a worthless fellow : **b: CHAP, FELLOW** <a gay ~>  

3a: any of various usu. simple mechanical devices for holding, gripping, or fastening consisting of a spike, rod or bar  

3b: **ANDRION**  

4a: **SUN DOG**  

4b: **WATER DOG**  

4c: **FOGBOW**  

**dog-like** \do\n. **adj**

---

**Entry:** dog  

**Pronounced as:** \do\n
**This entry defines the word when used as:** *n.*  

**The word occurs in** [ME]  

**The word is derived from** [OE docga]  

**Senses**  

1a: a highly variable carnivorous domesticated mammal (*Canis familiaris*) prob. Descended from the common wolf; broadly : any animal of the dog family (*Canidae*) to which this mammel belongs  

b: a male dog  

2a: a worthless fellow :  

b: **CHAP, FELLOW** <a gay ~>  

3a: any of various usu. simple mechanical devices for holding, gripping, or fastening consisting of a spike, rod or bar  

3b: **ANDRION**  

4a: **SUN DOG**  

4b: **WATER DOG**  

4c: **FOGBOW**  

**dog-like** \do\n. **adj**

---

**SPACE!**
Online dictionary **MACRO STRUCTURE**: Requires a well thought of lexicon structure!

Lexicographer needs to decide on the structure

1. Printed version CAN be a starter
2. Additional:
   - example sentences
   - encyclopedic information
   - multimedia
   - relational linking
3. Customization:
   - define different views for different users
Some history:

2006: Do we (&you) need a lexicon tool?

Yes:
(1) standardization for interoperability between lexica
data exchange
archive exploitation.
(2) community needs to participate in the construction of lexica
LEXUS

Based on two ISO TC 37 standards for linguistic resources

**LMF**: Linguistic 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

**Web based (!)**:
- Users work in LEXUS workspaces
- Workspaces can be shared by users with different rights (read-only, editor)
- Workspaces can contain multiple lexica
Show lexus lexicon first!
STRUCTURE and CONTENT

A lexicon needs a lexicon structure:

1. elements of the lexical entries
2. the hierarchy between them
3. the value domains
Content of the lexicon is built on the structure:
LEXUS: the lexical space

- Creation of lexica from scratch, import lexica from other formats (Toolbox, XML, Chat)
LEXUS: the lexical space

- Creation of lexica from scratch, import lexica from other formats (Toolbox, XML, Chat)
- User can customize views of word list and lexical entries
- Linking multimedia fragments to lexical entries

Audio
Video
Image
Link to archived files
LEXUS: the lexical space

- 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
• 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)
• 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)
• Search all lexica in the workspace or
• Setting filters on the word list of one lexicon
LEXUS: the lexical spaces

- Creation of lexica from scratch, import lexica from other formats (Toolbox, XML)
- 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)
- Search all lexica in the workspace or
- Setting filters on the word list of one lexicon
- Access rights (read/write) for registered users
ViCoS: the conceptual space

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)
User definable relation types

ViCoS functionalities – conceptual spaces
Default ‘Universal’ relation types

ViCoS functionalities – conceptual spaces

Relation Types

- is_a_kind_of
- is_not_a_kind_of
- is_part_of
- is_anonym_of
- is_synonym_of
- is_related_to
- is_caused_by

Relation Type Definition for: is_a_kind_of. To delete, press here: delete!

Name of relation type: is_a_kind_of

Import relation type from DCR - ISO 12820

Describe new relation type

The relation A is_a_kind_of B specifies that A has as hypernym B, or that B is a...

Example: car is_a_kind_of vehicle

This relation is a universal relation and can thus be applied across lexical resour...
ViCoS demo
The Iwaidja language of the Cobourg Peninsula, Northern Territory, Australia (approximately 200 speakers)

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
Iwaidja documentation project

| 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
Getting Toolbox into LEXUS

Structure comes from *.typ file

Content comes form *.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
\ee Unlike most named dugong cuts, there is no equivalent human part for this one.
\rf IwNo9Ma05
\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.
Getting Toolbox into LEXUS

Wara warrkbi abangkun.

There’s a big, tall man.

This contrasts with walarra 'tall and skinny, gangly'.

visible

Kudnayanjing janad rimardyarrwun ba wara, abangkun ba wara, riki bani marti.

You see that bloke there walking, that’s his father, this one here is his son.

Waribuyin kayidaruny ba kurlijuk, abangkun rajarrika kurlijuk.

(The mother emu) lays the eggs and leaves them, the father emu looks after the eggs.

11/Ful/2007
<table>
<thead>
<tr>
<th>x</th>
<th>alabanja</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>alabanja</td>
</tr>
<tr>
<td>sn</td>
<td>1</td>
</tr>
<tr>
<td>ps</td>
<td>2</td>
</tr>
<tr>
<td>de</td>
<td>beach hibiscus. Rope for harpoons and tying up canoes is made from this tree species, and the timber is used to make</td>
</tr>
<tr>
<td>sc</td>
<td>hibiscus</td>
</tr>
<tr>
<td>ge</td>
<td>hibiscus</td>
</tr>
<tr>
<td>re</td>
<td>hibiscus, beach</td>
</tr>
<tr>
<td>fs</td>
<td>205,410; IE 84</td>
</tr>
<tr>
<td>sd</td>
<td>plant</td>
</tr>
<tr>
<td>nd</td>
<td>material</td>
</tr>
<tr>
<td>rf</td>
<td>Iwa05.Feb2</td>
</tr>
<tr>
<td>wv</td>
<td></td>
</tr>
<tr>
<td>xe</td>
<td>hibiscus string/rope</td>
</tr>
<tr>
<td>sn</td>
<td>2</td>
</tr>
<tr>
<td>ps</td>
<td>2</td>
</tr>
<tr>
<td>de</td>
<td>short-finned batfish</td>
</tr>
<tr>
<td>ge</td>
<td>short-finned batfish</td>
</tr>
<tr>
<td>re</td>
<td>batfish, short-finned</td>
</tr>
<tr>
<td>sc</td>
<td><em>Zabidius novaemaculatus</em></td>
</tr>
<tr>
<td>sd</td>
<td>animal</td>
</tr>
<tr>
<td>sd</td>
<td>fish</td>
</tr>
<tr>
<td>rf</td>
<td>Iwaidja Fish Names.xls</td>
</tr>
<tr>
<td>sc</td>
<td>MELP project elicitation</td>
</tr>
<tr>
<td>eb</td>
<td>SH</td>
</tr>
<tr>
<td>dt</td>
<td>19/Dec/2006</td>
</tr>
</tbody>
</table>

Language consultants identify other fish species which have a moon-like shape as alabanja (e.g. juvenile of diamond trevally)
Getting Toolbox into LEXUS

Solution: chunking!
Getting Toolbox into LEXUS

Solution: chunking!
Contact:

Jacquelijn.Ringersma@mpi.nl for LEXUS user account
LEXUS and ViCoS support
Special functionality requests