

LEXUS and ViCoS

Jacquelijn 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 multi media encyclopedic dictionaries and lexica

ViCoS: a web based tool for constructing and visualizing conceptual spaces. Enrichment of lexical information with conceptual knowledge





LEXUS & ViCoS: From lexical to conceptual spaces



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 multi media lexicon (Steve Levinson)

Multimedia lexicon



Typed relations within the lexicon





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: 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

	THE THE WORDS W	ew search wellstion-types Help	niclas	
		Malayali: W	lord list	
Results 4-28 of 1488 for no e	ukir:			
Lexical Entry	Results 1-26 of 268 for at ent	ries		
3	A References Page 1	next Results 1-25 of 246 fer all entries	-	
2	6 #	THE PARTY REPORT	LEMMA	
1	e warat	8	АББЬР пімр 3 П	
N FER	d ostes	AAA 1	abb7	
and a set	e back (of pera an)	abbpah	MEDIA	
CITY CALLER	f and	attens	A REAL PROPERTY AND INCOME.	
new mans -	B Dark of Deep	attess	CONTRACTOR AND AND A	
ATAN.	h bely	ALTERATION	A REAL PROPERTY AND A REAL	
rgedternig	i _ 24	sitewows	C C C C C C C C C C C C C C C C C C C	
ngeterrag	J we	attes	A REAL PROPERTY OF THE REAL PR	
Hardikin	k bis	Barrent .	a line and a line of the second	
reaser	1	attaction of a	a strategy we have a strategy	
C. C	1000d	utres a		
	i bi swi (ci wind)		000000000	
			4 Contractor (1997)	
		appoint	TRANSLATION	
		altone	дождь м	



2008-2009:

Further refinements of ViCoS New LEXUS user interface





LEXUS development

2009: Launch of the tools, Workshop April 2, 2009

2006 \rightarrow 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



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 build 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







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 \rightarrow this facilitates searching cross lexica, sharing & merging

Archiving exploitation framework, XML based









Archive

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



Archive

- 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

naed access v conference v second conference v	Seeks of a ware a facebook finding View				Audio
lankul méten Kenar Kenar Jari Malak I Kepel, prochuji tari bir taji tari Malak I Kepel Kenaraki tarih Mala populati tal	farfar'üzə n role menog devole de man mente ray: devil ray	Constant of the Second	() (dau-10194 () (dau-10194	(until later) a set stay set	Video Image
A CALLY Score Instant in 1 respective particulution 2 and provide comparison particulution 2 and provide comparison particulution and in comparison particulution and in comparison particulution and an exercise of the comparison and an exercise of the comparison and an exercise of the comparison and an exercise of the comparison		economies - me - conomies -		range valorin v Idensid wind L Reniedad Romania Romania Romania Romania Romania Romania Romania Romania	Link to archived files
INTERPORTUNATION CONTRACTOR CONTRACTOR	februe neutóbles-tipg Aristanie g	n d real real contractions real		Landskeise someweite Merisgissen Merisgis	
a aamii aa		nik (stroker stroker ser te		Servery Contar	



- 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

Vei Drye Lexitor			
patron Providence (1994)	fre Sart order editor		
errie I			
veida			
acception:			
Start letter	Characters	Add	
	and all All net following (net following on the second sec	Remaine	
6	ps[-rX-eXrXe]		
4	45(-4)(-0)(~4)(~0)	- thear all	
0	(i)		
	P.*B.*B.*B.*B.*B.		
	$\left[h(\mathbf{x}_1^* - \mathbf{x}_2^* + \mathbf{x}_1^*) - h(\mathbf{x}_1^* - \mathbf{x}_2^*) \right]$		
	and the second se		
	04.01-410-01-41		
a	[witesit-witewitesitesitesitesitesitesitesitesitesites		
a 11	unal-uult-all-uult-all-all-all-all-all-all-all-all-all-a		
6 n	wa(-wf(-aff.caff.caff.caff.caff.caff.caff.caff.		
a 1 1	(+01(A01(+01)+A01(+A01(+A01)+A01) (A01(+01(+01)+A01(+A01(+A01)+A01)+A01) (A01(+01(+01)+A01(+A01(+A01)+A01)+A01) (A01(+01(+A01(+A01)+A01)+A01)+A01) (A01(+A01(+A01(+A01)+A01)+A01)+A01) (A01(+A01(+A01(+A01)+A01)+A01)+A01) (A01(+A01(+A01(+A01)+A01)+A01)+A01) (A01(+A01(+A01(+A01)+A01)+A01)+A01) (A01(+A01(+A01(+A01)+A01)+A01)+A01)+A01) (A01(+A01(+A01(+A01)+A01)+A01)+A01)+A01) (A01(+A01(+A01(+A01)+A01)+A01)+A01)+A01) (A01(+A01(+A01(+A01)+A01)+A01)+A01)+A01)+A01)+A01)+A01)		
4 7 7	w(-v)[-w][-w][-w][-w]] w(-v)[-w][-w][-w][-w]] w(-v)[-w][-w][-w][-w]]		
4 7 7 7	(d)((0)[-4][-4][-4][-4][-4][-4][-4][-4][-4][-4]		
1 m ns - -	(d)(d)[-d][-d][-d][-d][-d][-d][-d][-d][-d][-d]		
1 m ns - - -	(a)((a)((a)((a)((a)((a)((a)((a)((a)((a)		



- 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

SaltuDic Aug08 ride Booles s with	
hour have been by search	
NULARS	
ethiquette (energi com)	
el Ner « vazza»	
и Пер-чиска	
esca and a Road	
n autory	
NUMBER OF THE OWNER	
anna nyan	
niarim thus says short.	
DERIVATION	

Laxicon adite	
	7e be specified
	ерсик ник тих-Лфени срфифседник
	To be itself ef



- 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



Archive

Archive

- 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



mirdik

arrito

new guinea mudskipper bird species, seagui



- Adjustable display

ViCoS

Archive

- 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
 - \rightarrow 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



c	Marrgu	Iwaidja	ligar	Tiwi	Kunwinjku
'eye'	dhaida	yarta	yarta	pijara	kun-mim
'eyebrow'	bunurri	umburnuri	umbumumi	wumuringa	kun-milhngalk
'elbow'	waharti	waharti	waharti	jimpunga	kun-ney
(his) foot'	hi-rlud	ruli	yi-rkluli	kintanga	kun-denge
'moon'	mana	gurrana	hali	japarra	dird
'long ago'	larrud	wulamud	wularrud	parlingarri	korroko
'water'	wubaj	wubaj	wubaj	kukkuni	kukku

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 crossreferenced conceptual spaces will contribute to conservation, revitalization and empowerment of the language and community



Archive



Archive

Structure comes from *.typ file

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

*.typ file consists of marker definitions

\+mkr ps
\nam part of speech
\Ing Default
\mkrOverThis Ix
\-mkr

\+mkr sn
\nam sense number
\Ing Default
\mkrOverThis ps
\-mkr

Archive

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	<pre>\lx abarrk \lc 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</pre>
	\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



Archive

6 . 3 9	$ \begin{array}{c} \bullet \bullet$
\bx	blabania
Vc	alabanja
\sn	1
\ps	n
Ine	beach hibiscus. Rope for harpoons and tying up canoes is made from this tree species, and the timber is used to make fv(la
\sc	Hibiscus tiliaceus
\ge	hibiscus
/re	hibiscus, beach
. Vrfs	205,410, 正 84
\sd	plant
\sd	material
· · · · \/f	Iwa05.Feb2
þev	alabanja alhurdu
\xe	hibiscus string/rope
\sn	2
\ps	n
\de	short-finned batfish
\ge	short-finned batfish
\re	batfish, short-finned
\sc	Zabidius novaemaculatus
\sd	animal
\sd	fish
\rf	Iwaidja Fish Names.xls
\so	MELP project elicitation
. \eb	SH
. \dt	19/Dec/2006
\ch	Language consultants identify other fish species which have a moon-like shape as alabania (e.g. invenile of diamond trevally

Archive



1 : \lx\lc\a \ps\ge\de\xv\xe\rf\dt 1 : \lx\lc\a \ps\ng\sn\de\xv\xe\rf\ch\sn\de\sn\de\xv\xe\rf\dt 1 : \lx\lc\a \ps\pl\de\ge\re\ph\rfs\ng\nq\sd\rf\xv\xe\dt 1 : \lx\lc\a \ps\sn\de\ge\a \re\re\ng\sd\rf\xv\xe\rf\xv\xe\rf\xv\xe\rf\xv\xe\rf\xv\xe\sn\de\xv\xe\sn\de\rf\xv\xe\pd\ee\ 1 : \lx\lc\a \ps\sn\de\ge\ge\ge\rf\xv\xe\sn\de\ge\re\dt\pd 1 : \lx\lc\a \ps\sn\de\ge\ge\re\sn\de\rf\zv\xe\sn\de\ge\re\3s\ph\a \a \sd\dt 1 : \lx\lc\a \ps\sn\de\ge\re\re\re\re\rf\xv\xe 1 : \lx\lc\a \ps\sn\de\ge\re\ng\sd\rf\xv\xe 1 : \lx\lc\a \ps\sn\de\ge\re\ng\sd\rf\xv\xe\dt\sn\de\xv\xe\rf\pd\sn\ps\de\re\pd\2D\so\eb\dt 1 : \lx\lc\a \ps\sn\de\ge\re\re\re\rf\xv\xe\rf\x 1 : \lx\lc\a \ps\sn\de\ge\re\rf\xv\xe\rf\xv\xe\rf\xv\xe\rf\xv\xe\sn\de\ge\re\rf\xv\xe\rf\xv\xe\rf\xv\xe\sn\de\ge\re\rf\xv\xe\rf\xv\xe\sn\de\ge\re\rf\xv\xe\rf\xv\xe\sn\de\ge\re\rf\xv\xe\rf\xv\xe\sn\de\ge\re\rf\xv\xe\rf\xv\xe\sn\de\ge\rf\xv\xe\sn\de\ge\rf\xv\xe\sn\de\sn\de\sn\de\sn\de\sn\de\sn\de\sn\de\sn\de\sn\de\sn\sn\de\sn\de\sn\de\sn\sn\de\ 1 : \lx\lc\a \ps\sn\de\qe\re\re\sd\sd\xv\xe\rf\ng\sn\de\qe\re\xv\xe\rf\sn\de\qe\re\rfs\ng\sn\de\qe\re\rkxv\xe\rf\sn 1 : \lx\lc\a \ps\sn\de\ge\re\re\sn\de\ge\re\re\re\rfs\dt\ng\pd 1 : \lx\lc\a \ps\sn\de\ge\re\sn\de\ge\re\sd\rf\xv\xe\dt\pd 1 : \lx\lc\a \ps\sn\de\ge\re\sn\de\ge\re\sd\rf\xv\xe\ng\sn\de\ge\re\sd\sn\de\sd\so\dt 1 : \lx\lc\a \ps\sn\de\ge\re\rfs\cf\ng\rf\xv\xe\dt\rf\xv\xe\sn\de\re\ge\re\re\xv\xe\rf\xv\xe\rf\pd 1 : \lx\lc\a \ps\sn\de\ge\re\rf\xv\xe\sn\de\ge\ng\dt\pd 1 : \lx\lc\a \ps\sn\de\ge\re\rf\xv\xe\sn\de\ge\re\rfs\ng\dt 1 : \lx\lc\a \ps\sn\de\ge\re\rf\xv\xe\sn\de\re\re\ng\pd\rfs\dt 1 : \lx\lc\a \ps\sn\de\ge\re\sd\rf\xv\xe\dt\sn\de\ge\re\rf\xv\xe\rfs\xv\xe\pd 1 : \lx\lc\a \ps\sn\de\ge\re\sd\rf\xv\xe\sn\de\re\ge\sd\ng\dt 1 : \lx\lc\a \ps\sn\de\qe\re\sd<u>\sd\sn\de</u>\qe\re\pl\sd\sd\dt\eb 1 : \lx\lc\a \ps\sn\de\qe\re\sn\de\qe\cf\dt 1 : \lx\lc\a \ps\sn\de\ge\re\sn\de\ge\ge\re\re\sn\de\ge\re\rf\xv\xe\ng\sd\sd\rf\xv\xe\dt\pd 1 : \lx\lc\a \ps\sn\de\ge\re\sn\de\ge\re\rf\xv\xe\sn\de\rf\xv\xe\dt\pd\cf\ng\sd\ps\de\re\ge\pd 1 : \lx\lc\a \ps\sn\de\qe\re\sn\de\qe\re\sn\de\qe\re\rf\xv\xe\sn\de\qe\re\sn\de\qe\re\rf\xv\xe\dt 1 : \lx\lc\a \ps\sn\de\ge\re\sn\de\re\ge\ng\dt 1 : \lx\lc\a \ps\sn\de\ge\re 1 : \lx\lc\a \ps\sn\de\ge\re\xv\xe\ph\sn\de\re\rf\xv\xe\sd\xv\xe\rf\dt\eb\nq\dt 1 : \lx\lc\a \ps\sn\de\qe\re\xv\xe\xv\xe\rf\dt\sn\de\qe\rf\xv\xe\nq\dt 1 : \lx\lc\a \ps\sn\de\ge\sn\de\re\rf\xv\xe\rf\ps\sn\de\sn\de\ge\re\re\rf\s\xv\xe\rf\pd\dt 1 : \lx\lc\a \ps\sn\de\nq\qe\re\rf\xv\xe\rf\xv\xe\rf\xv\xe\sn\de\re\qe\rf\xv\xe\nq\dt 1 : \lx\lc\a \ps\sn\de\re\qe\sn\de\re\re\re\re\re\re\re\sn\de\qe\re\re\sn\de\qe\re\rfs\dt 1 : \lx\lc\a \ps\sn\de\re\re\qe\sn\de\re\qe\pd\dt 1 : \lx\lc\a \ps\sn\de\re\re\sn\de\re\re\sn\de\rf\xv\xe\pd\dt 1 : \lx\lc\a \ps\sn\de\rf\xv\xe\ng\ge\sn\de\sn\de\uv\rf\xv\xe\rf\xv\xe\sd\ng\dt 1 : \lx\lc\a \ps\sn\de\rf\xv\xe\rf\xv\xe\rf\xv\xe\sn\de\qe\re\sn\de\qe\re\rf\xv\xe\nq\dt 1 : \lx\lc\a \ps\sn\de\rf\xv\xe\sn\de\rf\xv\xe\sn\de\qe\ng\ng\sn\de\xv\xe\rf\xv\xe\rf\xv\xe\rf\so\eb\dt 1 : \lx\lc\a \ps\sn\de\rf\xv\xe\sn\de\sn\ps\de\xv\xe\ge\ng\dt 1 : \lx\lc\a \ps\sn\de\sn\de\xv\xe\ge\ge\ge\ge\ge\a \sn\de\rf\xv\xe\dt\pa\ng\rfs 1 : \lx\lc\a \sn\ps\de\ge\ng\rf\xv\xe\rf\xv\xe\sn\ge\re\ge\de\rf\xv\xe\rf\xv\xe\rf\xv\xe\sn\ps\de\dt 1 : \lx\lc\a \sn\ps\de\qe\re\dt\xv\xe\rf\sd\sn\ps\de\xv\xe\rf\sd 1 : \lx\lc\a \sn\ps\de\ge\re\re\re\re\re\re\rf\ng\sd\xv\xe\rf\dt\sn\ps\sn\de\ge\re\re\re\re\ng\sn\de\re\rf\sd\sd\dt\

🛦 Language

Archive



1 : \lx\lc\a \sn\ps\de\xv\xe\rf\sn\ps\de\xv\xe\sd\sd\dt





Show structure





