Ensuring Semantic Interoperability on Lexical Resources

M. Kemps-Snijders, C. Zinn, J. Ringersma, M. Windhouwer

Max Planck Institute for Psycholinguistics
P.O. Box 310, 6500 AH Nijmegen, The Netherlands
Marc.Kemps-Snijders@mpi.nl, Claus.Zinn@mpi.nl, Jacquelijn.Ringersma@mpi.nl, Menzo.Windhouwer@mpi.nl

Abstract

In this paper, we describe a unifying approach to tackle data heterogeneity issues for lexica and related resources. We present LEXUS, our software that implements the Lexical Markup Framework (LMF) to uniformly describe and manage lexica of different structures. LEXUS also makes use of a central Data Category Registry (DCR) to address terminological issues with regard to linguistic concepts as well as the handling of working and object languages. Finally, we report on ViCoS, a LEXUS extension, providing support for the definition of arbitrary semantic relations between lexical entries or parts thereof.

1. Introduction

The Max Planck Institute in Nijmegen hosts a 25 terabyte archive of multimedia linguistic resources including audio, video, and annotated multimedia files as well as text corpora and lexica `[corpus1.mpi.nl/ds/imdi_browser]`. These resources originate from past and current research projects that are addressing different scientific topics from various angles following different theories and approaches. Consequently, there is a large variation in the structure and terminology of research data. Also data is stored in different formats and character encodings. The diversity of data structures, formats and encodings makes it hard for researchers to explore and analyze data sets of others, in particular, when the amount of data requires computer-supported analysis. In this paper, we describe a unifying approach to tackle data heterogeneity issues for lexica and related resources.

2. LMF and DCR

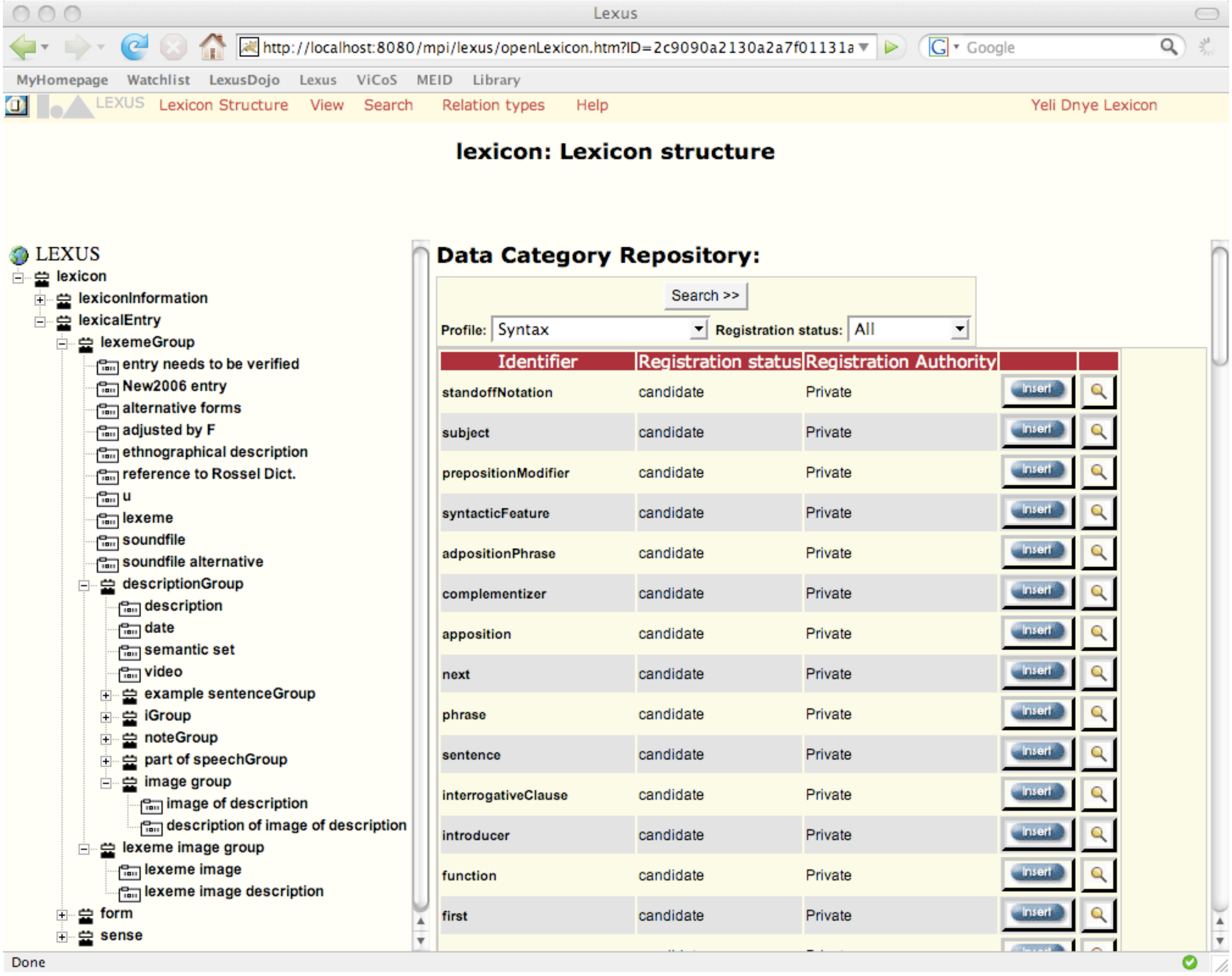
The Lexical Markup Framework Model (LMF) provides a simple core model with a flexible extension mechanism, thus capable of handling the structural differences that are encountered (Francopoulo et al., 2006). The LMF model also addresses terminological issues by incorporating the notion of data categories selected from a central Data Category Registry (DCR) (Wright, 2000). The DCR provides a shallow two-level hierarchy describing linguistic concepts and their associated value domain, if applicable, as well as a mechanism for handling working and object languages. For example, consider the data category `/grammaticalGender/` with its value domain containing `/masculine/`, `/feminine/` and `/neuter/`. For the French working language, a French definition may be provided; for the French object language, the data category `/grammaticalGender/` may be restricted to only `/masculine/` and `/feminine/`. Users may select data categories from the DCR and embed them into the LMF model to create schemas reflecting the requirements for their scientific work.

3. LEXUS

The LEXUS tool is based on the work of LMF and allows users to create, read and modify lexicon structures (Kemps-Snijders et al., 2006). The core model consists of a `Lexicon`, containing `LexicalEntry`s. Each `LexicalEntry` consists of a `Form` and `Sense` pair. Lexical entries are thus form-sense oriented. The core model may be extended by adorning it with data categories selected from the DCR or from the Shoebox MDF format (representing a list of linguistic concepts commonly used by our field linguists). It is also possible to create user-defined data categories with no reference to any standardized linguistic concept. Here, users define their own scientific notation. Fig. 1 shows a screenshot of LEXUS after a user chose to define a new data category with the help of the DCR. To further assist users in the creation process of their lexicon schema they may also make use of the Component Registry. This registry hosts a number of predefined sub-schemas that may be plugged into an LMF schema. These sub-schemas represent best practices schemas (*e.g.*, for Morphology) for given linguistic theories or specific research objectives. The construction of schemas thus becomes much easier and thus also enables less experienced users in defining lexicon structures.

Lexical entries created with LEXUS will adhere to the specified schema. LEXUS also supports the import of lexicon files if they are in one of the following formats: XML, Shoebox and Clan. During the import process, parts of the structures may be rearranged to better reflect the intentions of the LMF model. The unified model approach thus allows for search and comparison across heterogeneously structured lexica stemming from different sources. The character encoding in which all data is stored is UTF-8.

When it comes to data entry, LEXUS provides multimedia support. Multimedia fragments provide the possibility to supplement lexica with much richer information than is possible in traditional lexica. Sound files can be used, for instance, to illustrate pronunciation, while video and annotated media may be used to provide examples of language use in real life settings. With LEXUS/LMF, audio, video and other multimedia resources can be attached to any level in a lexicon structure. It is also possible to link to archived annotated media file segments, given the availability of archive integration and supporting tools. These will be displayed using ANNEX, the annotation exploration tool built at the MPI (Berck and Russel, 2006).



lexicon: Lexicon structure

- LEXUS
  - lexicon
    - lexiconInformation
    - lexicalEntry
      - lexemeGroup
        - entry needs to be verified
        - New2006 entry
        - alternative forms
        - adjusted by F
        - ethnographical description
        - reference to Rossel Dict.
        - u
        - lexeme
        - soundfile
        - soundfile alternative
        - descriptionGroup
          - description
          - date
          - semantic set
          - video
          - example sentenceGroup
          - iGroup
          - noteGroup
          - part of speechGroup
          - image group
            - image of description
            - description of image of description
        - lexeme image group
          - lexeme image
          - lexeme image description
    - form
    - sense

Data Category Repository:

Profile: Syntax Registration status: All

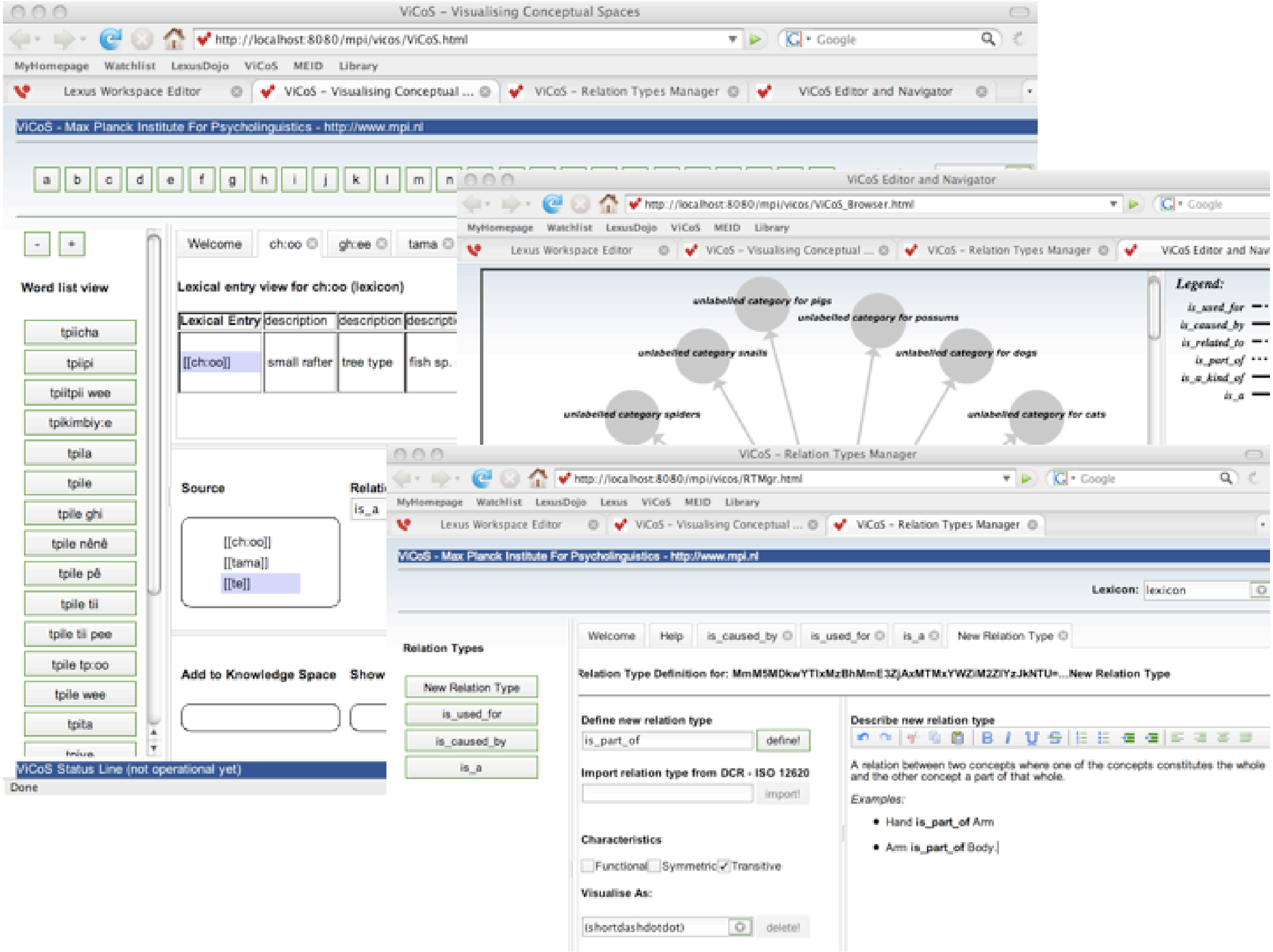
Identifier	Registration status	Registration Authority		
standoffNotation	candidate	Private	Insert	🔍
subject	candidate	Private	Insert	🔍
prepositionModifier	candidate	Private	Insert	🔍
syntacticFeature	candidate	Private	Insert	🔍
adpositionPhrase	candidate	Private	Insert	🔍
complementizer	candidate	Private	Insert	🔍
apposition	candidate	Private	Insert	🔍
next	candidate	Private	Insert	🔍
phrase	candidate	Private	Insert	🔍
sentence	candidate	Private	Insert	🔍
interrogativeClause	candidate	Private	Insert	🔍
introducer	candidate	Private	Insert	🔍
function	candidate	Private	Insert	🔍
first	candidate	Private	Insert	🔍

Figure 1: LEXUS' interface to the DCR. On the left the schema definition of the Yélî Dnye lexicon, on the right data categories from the Syntax Profile of the DCR.

During the creation process of a lexicon, the user may also define lexical entry views. For this, users can use a simple built-in editor, or make use of an external HTML editor of their choice, to design a template that defines how the information stored in a lexical entry is to be displayed. Multiple views can be defined. When displaying a lexical entry in a word list, only selected pieces can be shown on a single line; while a full lexical entry view, having no space restrictions, may show all available information as well as attached multimedia resources. In addition, word list views will take a user-definable sort order into account. In fact, users may define a custom sort order for each of the data categories that appear in their lexicon, and access to the lexicon is provided through any element specified in the sort order. Characters in the sort order may be any UTF-8 character, thus allowing access to the lexicon using, for example, IPA characters or (varieties of) Cyrillic characters.

LEXUS also offers base support for the creation of relations between lexical entries or parts thereof, or even across lexica. A lexical resource can thus become more than a flat list of lexical entries that adhere to the same structure; lexical resources with relations have a second layer of structure that can be used to systematically encode linguistic information, using lexical entry information, but not necessarily storing it there. Relations can also be used to address semantic heterogeneity issues between lexica, and to support, for instance, the merging of lexical resources. We start with intra-lexica relational linking.

In the past, researchers have used existing LEXUS functionality to encode and maintain ontological information that is rooted in the words to express them. We give an example. The lexicon for the Yélî Dnye language (which is spoken on Rossel Island, Papua New Guinea) currently contains over 6000 lexical entries; mostly nouns representing objects and entities in the natural world. This lexicon has being created with Toolbox [www.sil.org/computing/toolbox] during and following field research, and later has been imported into LEXUS. One objective is to use its linguistic information to (re)construct a conceptual space that represents the natural world from an ethnobiological perspective. In fact, no systematic ethnobiology has been done on Rossel Island hence little biological taxa has been established. The information in the Yélî Dnye lexicon, however, hints at a great richness of tradi-


ViCoS - Visualising Conceptual Spaces
http://localhost:8080/mpi/vicos/ViCoS.html
MyHomepage Watchlist LexusDojo ViCoS MEID Library
Lexus Workspace Editor
ViCoS - Visualising Conceptual ...
ViCoS - Relation Types Manager
ViCoS Editor and Navigator
ViCoS - Max Planck Institute For Psycholinguistics - http://www.mpi.nl
a b c d e f g h i j k l m n
- +
Word list view
tpiicha
tpiipi
tpiitpii wee
tpikimbiy:e
tpila
tpile
tpile ghi
tpile nênê
tpile pê
tpile tii
tpile tii pee
tpile tp:oo
tpile wee
tpita
Welcome
ch:oo
gh:ee
tama
Lexical entry view for ch:oo (lexicon)
Source
Relati
is_a
[[ch:oo]]
[[tama]]
[[te]]
Add to Knowledge Space
Show
ViCoS Status Line (not operational yet)
Done
ViCoS Editor and Navigator
http://localhost:8080/mpi/vicos/ViCoS_Browser.html
unlabelled category for pigs
unlabelled category for possums
unlabelled category snails
unlabelled category for dogs
unlabelled category spiders
unlabelled category for cats
Legend:
is_used_for
is_caused_by
is_related_to
is_part_of
is_a_kind_of
is_a
ViCoS - Relation Types Manager
http://localhost:8080/mpi/vicos/RTMgr.html
MyHomepage Watchlist LexusDojo Lexus ViCoS MEID Library
Lexicon: lexicon
Relation Types
New Relation Type
is_used_for
is_caused_by
is_a
Welcome
Help
is_caused_by
is_used_for
is_a
New Relation Type
Relation Type Definition for: MmM5MDkwYTlxMzBhMmE3ZjAxMTMxYWZiM2ZlYzJkNTU=...New Relation Type
Define new relation type
define!
Import relation type from DCR - ISO 12620
import!
Characteristics
Visualise As:
delete!
Describe new relation type
Google

Lexical Entry	description	description	descripti
[[ch:oo]]	small rafter	tree type	fish sp.

Define new relation type: is_part_of

Characteristics: □ Functional □ Symmetric ☑ Transitive

Visualise As: (shortdashdotdot)

A relation between two concepts where one of the concepts constitutes the whole and the other concept a part of that whole.

Examples:

- Hand is_part_of Arm
- Arm is_part_of Body.

Figure 2: ViCoS GUIs for defining and browsing conceptual spaces, and for defining semantic relations.

tional knowledge about the biological environment that can be harvested. The lexical entries of the Yélî Dnye lexicon are organised by a hierarchical structure of data category groups (*e.g.*, `descriptionGroup`, `noteGroup`) and data categories (*e.g.*, `/lexeme/`, `/description/`, `/example/`, `/note/`) and their respective values. A LEXUS-supported full-text search query ”fish” on data category `/description/` will return the following two results, among others:

- **lexeme:** d:éél:a

  description: fish sp. (grouper)

- **lexeme:** doo

  description: chasing fish into a net

  example sentence: Kî néépi wunê doo.

  free translation of example sentence: That canoe is chasing the fish into the net.

In fact, the search query would contain most of the fish species of the lexicon (first result), but also some other lexical entries which are otherwise related to fish (second result). Some entries are missed though, namely those which only have a reference to ”fish” in one of their other data categories, for instance:

- **lexeme:** lete

  description: dolphin

  example sentence: lete te dyêêdî ngmê

  free translation of example sentence: the dolphin is a kind of fish

To also obtain such entries, a complex search query (OR search) across data categories can be performed. Once lexical entries with the required information have been localised, LEXUS offers basic support to construct and visualise relations between them.

From past usage, it is clear that the relation functionality will also induce semantic heterogeneity problems as users are free to define relation types as they see fit. In the given example, our researcher used ”is-a” to label its relation type, and added free text to describe it. Other researchers may prefer the label ”is a kind of” with identical meaning, and yet others may use an existing label but not as intended. This issue can be addressed with the help of the DCR. The DCR has increasingly rich resources with regard to standardised relation types. The terminology profile of ISO12620 proposes, for instance, relation types for specifying thesaurus-like relations that specify whether a term is broader than or narrower than another term

(/broaderTerm/, /narrowerTerm/), or for specifying that two concepts are regarded equal (/synonym/) or related via some association (/relatedConcept/). The semantic section of ISO12620 contains many more relations, most of which have registration status private rather than being standardised by a registration authority (*e.g.*, ISO/INRIA-LORIA, ISO/DublinCore). Semantic relations include /causes/ ("$1 causes/motivates/justifies $2"), /elaboration/ ("Binary relation which means that $2 is an elaboration of $1"), and /processStep/ ("$2 is step in process $1"). The definition of these relations, however, is given in free text, and DCR entries with private status often miss concise definitions. In time, these inaccuracies or flaws will be repaired, so that the DCR will serve as common reference point to address data heterogeneity issues.

4. ViCoS

The new ViCoS module extends existing LEXUS functionality on front-end and back-end (Zinn et al., 2008). On the interface side, it streamlines the process to define relations and relation types. Fig. 2 depicts elements of the ViCoS interface. The main window shows the word list view, a user-definable lexical entry view as well as the knowledge space pane. The form and content of all information stems from LEXUS. Elements of the lexical entry can be entered (via drag&drop) as source or target of a relation. Relation types can be selected from a set of pre-defined relation types, or users can define new relation types. For this, ViCoS' relation type manager gives access to the ISO-12620 data category registry. Moreover, ViCoS allows users to define relations that are specific to lexica. Such relations may be named using the language being documented, or deviate to some extent from the meaning of equivalent English-named relation types. While our current experience was restricted to intra-lexica relational linking, we believe that linking across lexica (which is already possible) will enable researchers to link together heterogeneous lexical resources within a single uniform framework. Fig. 2 also displays the ViCoS browser to navigate and further manipulate conceptual spaces. Continuing our example, researchers may look at fish species in lexica that describe other languages of Oceanian speech communities, say to study their variation in lexicalisation.

On the representation side, ViCoS augments the LEXUS representation format — a relation was stored in a proprietary database format — by the more expressive OWL knowledge representation language [http://www.w3.org/TR/owl-features/. Relation types and their instances are managed with the JENA framework [jena.sourceforge.net]. Consequently, ViCoS can then also offer reasoning services on OWL-based ontologies. Researchers may ask, for instance, whether a given speech community is regarding a dolphin as a type of fish, or query the OWL databases for birds that cannot fly. ViCoS will return a corresponding Conceptual Space which users can then explore or use as an entry point to the lexical space and attached multimedia resources.

5. Conclusion

In this paper, we described our approach for tackling data heterogeneity issues for lexica and related resources. Its central pillars are the Lexical Markup Framework together with the Data Registry Category, which help defining the structure and content of lexical resources. Lexica built with LEXUS, or imported by it, become resources that can be managed in a standardised way and thus become more accessible to a wider audience of researchers. With ViCoS, we offer a third pillar. Users can define their own relation types (when re-use of existing ones is impossible) and link together arbitrary lexical entries within or across lexica. This creates a semantic layer that can be used to better bridge any existing data heterogeneity gaps.

Notes

The website [http://www.lat-mpi.eu] serves as the Language Archiving Technology portal of the Max Planck Institute of Psycholinguistics. LEXUS and VICOS form part of the Language Archive Technology toolset, and more information on these and the other tools is available at this site.

6. References

P. Berck and A. Russel. 2006. Annex - a web-based framework for exploiting annotated media resources. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

G. Francopoulo, N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet, and C. Soria. 2006. Lexical markup framework: Iso standard for semantic information in nlp lexicons. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*. Also, see http://www.lexicalmarkupframework.org.

M. Kemps-Snijders, M-J. Nederhof, and P. Wittenburg. 2006. Lexus, a web-based tool for manipulating lexical resources. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

S. E. Wright. 2000. A global data category registry for interoperable language resources. Available at http://isotc.iso.org (ISO TC 37).

C. Zinn, G. Cablitz, J. Ringersma, M. Kemps-Snijders, and P. Wittenburg. 2008. Constructing knowledge spaces from linguistic resources. In *CIL 18 Workshop on Linguistic Studies of Ontology: From Lexical Semantics to Formal Ontologies and Back.*