Managing Complex and Multilingual Lexical Data with the Simple Editor

Preben Wik, Lars Nygaard and Ruth Vatvedt Fjeld
University of Oslo
Box 1001 Blindern, N-0315 Oslo
mail@prebenwik.com, lars.nygaard@ilf.uio.no, r.e.v.fjeld@inl.uio.no

Abstract
This paper presents an editor for compiling a multilingual machine readable lexicon, like the Simple-lexicon. This editor has proven to be a useful tool in linking several languages in one lexical database, and to edit the entries in a consistent and convenient way. The editor has been designed for linking Danish, Swedish and Norwegian in the Simple Scan-project, but might easily be extended to include all the languages in the Simple project. The editor may also be modified for similar machine readable lexical projects.

1. Introduction

Modern language technology and lexicography need lexical resources that are complex and multilingual. The Simple-lexicon is an attempt to develop of such a complex linguistic database (Lenci et al. 2000)

Simple-Scan is an ongoing project to link the Scandinavian Simple lexica together, using the Simple Editor, creating a basis for a trilingual dictionary (cf. Pedersen & al. 2002). A tool that could handle editing, creating, and browsing of such complex, multidimensional resources was created at the University of Oslo, and will be described in this paper.

2. The Simple Project

SIMPLE (Semantic Information for Multifunctional Plurilingual Lexica) is part of a European Union Language Engineering Programme. The aim of the project is to create a harmonised common model for encoding structured semantic information and compile descriptions of a core vocabulary of 10.000 word senses for 12 European Languages.

The central element in the Simple-lexicon, called a Semantic Unit (SemU), embeds a variety of other elements, some as features, some as links to other SemUs. Part of the features in the SIMPLE structure is the set of ontologies it holds. The underlying model for the development of the Core Ontology for SIMPLE (Simple-type) is the Generative Lexicon (Pustejovsky, 1998), allowing word senses to differ in terms of their internal complexity. Apart from the Simple-Type (which also contains "SuperType" and "UnificationPath"), two other ontologies are being represented: the "Semantic class" and "Domain" ontologies developed by Lexiquest.

SIMPLE also holds a set of Qualia relations and features. The qualia relations are grouped into four kinds: Formal, Agentive, Telic, and Constitutive, following the ideas of Pustejovsky's Generative Lexicon (Pustejovsky, 1998) The hyponym-hyperonym relation is for example part of the Formal group and encoded as "Isa" together with a link to another Semu. A large set of other relations such as "HasAsMember", "IsaPartOf", "Createdby", and "Isthehabitof" are contained in these four groups. In addition there are also Polysemic relations, Synonymy relations and Constitutive Features. Constitutive Features are not links to other Semus, but Features with values such as: Sex (Male, Female) or State (Solid, Liquid, Gas).
Another important aspects of SIMPLE is the specification of selectional restrictions for predicative semantic units. The task involves specifying the argument structure, and assigning a semantic marker (Simple-Type) to the arguments selected by a given verb, adjective or predicative noun.

3. The Simple Editor

The Simple Editor was written to accommodate the needs of two different kinds:

- A tool that enables lexicographers to edit entries in a consistent and convenient way, to quickly create new ones, and to create multilingual linking or compare semantic encoding across languages.
- A tool for Linguistic researchers and students to get information out of the SIMPLE-lexicon to be able to do a variety of searches, and create lists of words that hold some semantic relationship.
- A tool for error fnding and correcting the encodings in the database.

The editor consists of:

- a program that parses the SGML files that are used as an exchange format in the Simple Project, and inserts data in the database. (SGML is well suited for data transfer, but searching, browsing and editing complex SGML files is bound to be slow and error prone.)
- a graphical user interface to search, edit, and browse entries in the database**.**

3.1 Searching

Having transferred the data to a relational database allows for complex and efficient searches through SQL-queries. Users do not have to write SQL-queries themselves. Searches are performed by filling in one or more entry fields. The "_" and "%" signs can be used as wildcards in a search to widen the search possibilities.
When fields can only consist of entries from a closed list, the user is presented with these choices.

The ontologies are hierarchically structured, and this structure is also presented to the users. When searching for words tagged with a certain type, the search is also taking the hierarchy into account, so that the query contains the chosen type and all of its children. E.g. searching for "Animal" will also return SemUs tagged with "Earth-Animal", "Air-Animal", and "Water-Animal".

The results from a search is presented in a separate list. Normally the list will be emptied between each search, and hence show a list of words that corresponds exactly to the

restrictions of the search. But the results list can also be used to collect SemUs from a number of searches. This allows for the composition of lists of SemUs that holds some semantic relationship, for viewing, editing, or exporting purposes. This list can further be sorted in various ways. It is possible to export plain lists with just a selected number of features, or export the list back to the SGML exchange format. SemUs can also be deleted from the database through the results-list. Referential integrity is checked, and a warning will be issued if some other SemU is linked to the selected one through a qualia relation.

3.2 Browsing

Browse-mode demonstrates an entirely different way to view the data. Rather than viewing all the information available for each SemU one by one - as is done in the edit mode - chains of selected Qualia-relations are viewed simultaneously.
The BrowseTab displays parents several levels up, and children one level down of a selected qualia relation to a selected SemU. For the 'Isa' relation of the SemU 'Cat' It would look like: *'wildcat', 'house cat', 'kitten'...<Isa> 'Cat' <Isa> 'feline' <Isa> 'mammal'...*, but other qualia relations, appropriate for other SemUs such as Isin, Isapartof, or Hasasparts could also be selected, and present other chains of relations.

Some errors from the encoding process are easily discovered by viewing the data this way. Because the SemUs are seen in context, together with other SemUs that in a particular aspect are similar to a selected SemU, human inspection will quickly notice inconsistencies and SemUs that does not belong to the group can be re-linked.
Some examples might be in place to explain this.

- Consistency in choice of granularity
  Europa <Isin> Verden (The world)
  England <Isin> Europa
  Argentina <Isin> Verden

  Seeing it isolated there is nothing wrong with saying Argentina <is in> the world. Seen in context however, it is clear that the level South-America was bypassed.

- Circularity
  Sometimes it is not all clear which is a hyponym and which is the hyperonym, and circular definitions might occur.
  instutition<isa>foretagende<isa>virksomhed<isa>institution...
  Again, isolated it might be hard to detect errors that can become evident in the right context.

3.3 Editing

In editing-mode, a comprehensive list with all the information of the selected SemU is displayed. Users can also select a reference language, and a second list will be displayed, with a corresponding SemU in the selected reference language. This is useful for creating multilingual linking or for comparing the semantic information encoded across languages.

A selected SemU can be edited in various ways. Each field can be edited manually, and different dialog windows will appear depending on which type of field it is. Similar to

the Search-mode, users are presented with a list of choices on the fields where the field information is accessible in the form of a list, in order to avoid errors from wrong spelling.

The Glossa (freedefinition) field is a link to an online monolingual dictionary. If more than one word meaning exist for the chosen lemma, users are presented with a list where they can choose the appropriate definition.

When editing the Predicative representation field, a separate editor opens up, where users can search and select the appropriate argument structure and predicate elements.

When editing Qualia relations, a separate editor opens up, where users can select qualia type and search for the SemU to link with, ensuring that linking is only done to existing SemUs.

For convenience, entries can also be copied directly from the corresponding SemU in the reference language.

New SemUs can quickly be created by 'cloning' an existing SemU with similar characteristics.

3.4 Availability

The Simple Editor could easily support the other 12 languages in the Simple Project, thus facilitating the expansion of these lexica, and cross-linking of the entries in these languages - an integral part of the project (Villegas et.al, 2000). The program could also quite easily be modyfied to edit similar projects for lexical databases like WordNet and EuroWordNet.

The program is available under an open source licence (The Gnu General Public Licence), uses the open source database engine MySQL and runs under Windows, Unix and Mac OS X.

4. References

Lenci, A., N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruimy, M. Villegas and A. Zampolli. 2000. 'SIMPLE: A General Framework for the Development of Multilingual Lexicons', in International Journal of Lexicography, vol. 13, number 4, december 2000

Pedersen, Bolette, Ruth V. Fjeld and Maria T. Gronostaj. 2002. Harmonisering og sammenkædning af sprogteknologiske ordbaser med særligt henblik på informationssøgning - en rapport fra SPINN-netværket, in Holmboe, H (ed.) Nordisk Sprogteknologi/Nordic Language Technology. København.

Pustejovsky, J.1995. The Generative Lexicon, Cambridge, MA, The MIT Press. Villegas 2000