LEXUS, a web-based tool for manipulating lexical resources

Marc Kemps-Snijders, Mark-Jan Nederhof, Peter Wittenburg

Max-Planck-Institute for Psycholinguistics
Wundtlaan1, 6525 XD Nijmegen, The Netherlands
{marc.kemps-snijders,markjan.nederhof,peter.wittenburg}@mpi.nl

Abstract

LEXUS provides a flexible framework for the maintaining lexical structure and content. It is the first implementation of the Lexical Markup Framework model currently being developed at ISO TC37/SC4. Amongst its capabilities are the possibility to create lexicon structures, manipulate content and use of typed relations. Integration of well established Data Category Registries is supported to further promote interoperability by allowing access to well established linguistic concepts. Advanced linguistic functionality is offered to assist users in cross lexica operations such as search and comparison and merging of lexica. To enable use within various user groups the look and feel of each lexicon may be customized. In the near future more functionality will be added including integration with other tools accessing lexical content.

1. Introduction

At the MPI for Psycholinguistics and at many other places one is confronted with many different lexical structures and formats. In the pilot phase of the DOBES programme [1], where we had seven language documentation teams we were already confronted with twelve different structures. These were created with the help of a variety of tools such as Shoebox [2], relational database programs, EXCEL and WORD. This situation has evolved to the one now where we have 30 documentation teams in the DOBES programme and a comparable number of researchers at the MPI. In addition, we have large lexica such as from the CELEX [3] and the Dutch Spoken Corpus projects [4]. One reason for creating different formats lies in people's preferences for certain tools – they like to use those tools which they are used to, since this seems to guarantee efficiency. The reasons for different structures can be found in differences between studied languages, theories of the researchers and objectives of the projects. In particular, the latter can be very different. In the documentation of endangered languages, some teams argue that it is important to gather as many words as possible and to describe them in a shallow way, so that simple "table like" structures (headword, translation, morphological breakdown, examples) are already sufficient. As an example of a Natural Language Processing framework CELEX uses a very detailed structure containing about 35 related tables. These are necessary to be able to encode all envisaged attributes. An overview of the variety in structures was given by Wittenburg, Peters and Drude [5].

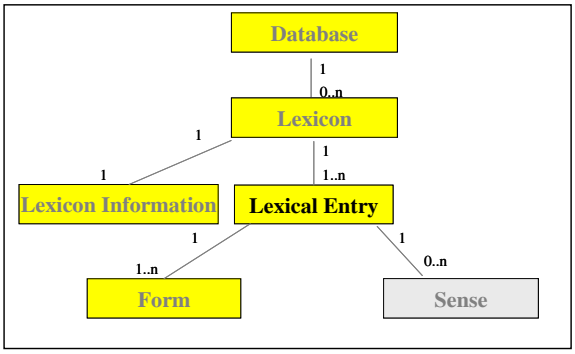
The huge variety makes it almost impossible for researchers to carry out cross-language operations, to integrate lexica with corpora etc. Users first would have to write converters, to learn to use different tools or to hire several student assistants, which is not feasible for most linguists given the high work load and the budget limitations. So, many interesting research questions will not be tackled. Also for software developers structural variety is a nightmare, since building new tools with new functionality would in the extreme case be limited to a certain format and structure. Therefore, people have looked for ways to overcome these barriers for some time. We can refer to the work of the Genelex project which defined an extensive schema for NLP lexica [6] or more recently the work within the ISLE/MILE project to work out structures for multilingual lexica [7]. We should not forget to mention the excellent Shoebox tool for field linguists which supports a wide range of lexical structures and that already provides researchers with a flexible framework to manipulate structure and content.

2. Lexical Markup Framework

The most recent and most promising approach to come to a generic framework for all sorts of lexica was started within ISO TC37/SC4 [8]. The Lexical Markup Framework (LMF) [9], currently being developed within ISO TC37/SC4, aims to become a standard which unifies lexical resources by providing a common model with a sufficient expressive power. By representing lexical resources in LMF comparing, exchanging and merging lexica, for example, will be simplified. It will allow linguists to have easy access to a wider range of lexical resources from various sources and can serve as a widely agreed and supported interchange format.

LMF consists of a core model supplemented by data categories, components and an extension mechanism for flexible creation of lexicon structures. The core model is kept very simple to not exclude concrete lexica (see figure 1) and it describes the mechanisms to extend this basic lexical form. Extensions using components and data categories may be added at any level, for example to create a morphological extension on a lexical entry or to add general information details to a lexicon. Information is aggregated into ComponentInstances and individual atomic linguistic information units (LIU). ComponentInstances refer to components while LIUs are associated with data categories. Data categories are properly defined linguistically meaningful concepts registered in a data category registry such as the one to be worked out in ISO TC37/SC4. The abstract definition of components include the possibility of typed relations, i.e., relations are components connecting LMF elements. Due to this definition relations can also have complex structure. By supporting selections of data categories

taken from widely accepted registries a platform for semantic interoperability between lexica is created.



The core model of LMF is kept very simple and therefore generic. It basically states that a lexicon has a number of lexical entries and some global information. Each lexical entry must have at least one form and it can have senses. To be able to represent relations between several lexica such as they may occur in multilingual lexica a database node was added. The core model also specifies how this core model can be extended.

The LMF standard is supplemented by a number of extensions that are typically found in different applications. These are not part of the core standard, but can be seen as examples and recommendations.

The MPI tested this model against all lexica they have in their repository and institutes like ILC in Pisa tested whether lexicon structures such as worked out in the SIMPLE [10] and PAROLE [11] projects can be represented. At present we are not aware of any serious omissions in the LMF specification and we expect that it will stabilize in 2006 and enter the formal standardization procedure.

3. LEXUS Tool

The LMF specification prompted the development of concrete tools to work with and test the model. LEXUS [12] is a new web-based framework that is fully compliant with the LMF model. It is the first implementation of LMF and demonstrates its usefulness and its possibilities. LEXUS provides a user-definable user interface to enable the creation and manipulation of arbitrary lexical structures. Components and data categories may be added, removed, altered or even rearranged by the user. LEXUS interacts with a number of Data Category Registries, in particular the one currently being established within ISO TC37/SC4 [13], to allow users to re-use widely accepted data categories from registries. Also the Shoebox MDF categories are supported since they are widely used by field workers. The system interacts with the ISO data category registry via an open API including the extraction of value sets. This data category selection process promotes semantic interoperability between lexica maintained by LEXUS. Additionally, users may define data categories of their own to describe linguistic concepts to support researchers who have to deal with different languages and who want to try out new theories, for example.

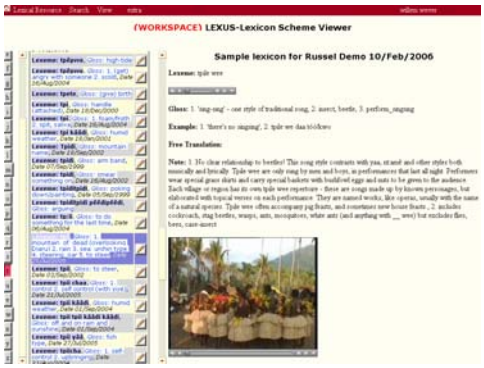
LEXUS also allows users to add and manipulate content. Content can be constrained by vocabularies and it is possible to include references to other resources. This allows LEXUS to integrate images, sound and video information. It requires the presence of appropriate players on the client to be able to visualize the different media. Flexible integration of media is particularly important for purposes of language documentation and for applications involving sign language applications.

Based on the lexical resource structure, the data input process is guided to ensure consistency between the data and the structure. On the one hand data input is restricted by the defined structure. On the other hand modifications to the structure are reflected in the data. Modification of lexical resources is done in the user's workspace ensuring that users can work independently of each other. However, this feature also allows multiple users to work on the same lexicon. Each authorized user may add or remove information to/from the lexicon or even change its structure. LEXUS is fully UTF-8 compliant, can export the chosen structure as an XML Schema and can export the lexicon itself as an XML file in TMF format [14].

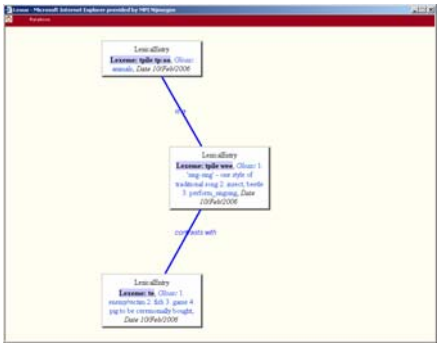
Due to the generic nature of LMF and the LEXUS implementation it should be useful for different user groups (LEXUS should be useful by language engineers as well as by field linguists). This can only be achieved when the user interfaces can be adapted to meet the various needs of these user groups.. The user interface has been re-designed three times to optimize acceptance. In the begin phase the user interface, for example, had a strong emphasis on the lexical structure and access to the content was a secondary step. We came to understand that many users do not approach lexica this way – structures are often implicit. We therefore had to modify the design to offer access to the content first and only show structure information on request. In addition, users who are willing to read the manual may customize many aspects of the user interface. In particular, the user can design a lay-out similar to those found in traditional written lexica. This allows users to also create printed lexica although some aspects of rendering of LMF lexica in printed form, such as media frame positioning, are not yet solved. The following figures give an impression of the current user interface look and feel.

A powerful feature of LEXUS is its possibility to draw typed relations and to visualize them as indicated in the following figure. This allows users to build up encyclopedic structures and to visualize them, as Manning experienced with his KirrKirr example [15]. This way of looking at a lexicon is much more preferred by, for example, language community members. To meet the goal for enabling format interoperability LEXUS offers import and export modules for Shoebox [16] and CHAT [17] lexica. For lexica written in Excel and Word there are converter options. However, experience shows that each lexicon has to be treated separately, since structure and

encoding is not consistent in general. Often, some pre-processing is required to filter out inconsistencies.



This figure shows a list of lexical entries in a form that can be determined by the researcher and it gives more information including multimedia streams for a selected entry. Also the way this information is presented can be defined by the user.



This figure shows a typical relational diagram where one word found in a lexical entry can have arbitrary relations with other words found in the same or even other lexica. By clicking on boxes the user can navigate for example through large semantic networks.

Common tasks in the use of lexical resources are lookup and comparison of lexical entries. LEXUS allows the user to specify complex search patterns and to search through multiple lexica at the same time. This search enables users to specify the linguistic concepts they are interested in and retrieve the relevant information even if the fields have different names. It is therefore possible for example to compare a “partofSpeech” data category in one lexical resource with a “POS” tagged data category in another resource. Yet, there is no support to allow users to create, store and re-use bottom-up driven mapping files make. This will be one of the next extensions.

4. Merging

By merging, the process of composing a single lexicon out of several lexica is meant. Electronic or printed lexica are often based on information stemming from different sources. In the context of the DOBES project for example, two or more field linguists may compose lexical entries on the same subject language independently from each other. Frequently, the resulting data may be structured differently and contain overlapping and possibly inconsistent information. The input lexica will usually involve a single subject language, but also merging of bilingual lexica are considered. Furthermore, the kind of information stored in the respective input lexica may be very different. One may contain syntactic information, whereas another contains semantic information, or geographical locations where certain words are used, etc.

Manual merging of lexica is a laborious and error-prone task. Especially when time and money are severely limited, automated support of this process is desirable. We have therefore designed a general model for the process of merging. This model incorporates tasks such as the identification of related lexical entries, restructuring of lexical information, and handling of inconsistent data. The model also offers much flexibility in deciding which tasks are to be done automatically and which are to be done manually. The linguist may monitor and influence every individual step in the creation of the new lexicon, and override all values that were derived automatically. At the other extreme, the user may choose to let LEXUS produce the entire target lexicon without any manual intervention.

The nature of the merging process depends foremost on the structures of the source lexica and on the structure of the target lexicon. It further relies on interaction with the user, who may indicate preferences to let certain tasks be done automatically, by means of specialized linguistic functions, or to have LEXUS return the control to the user at certain moments.

After the merging procedure has been established, the actual merging process can start. Lexical entries that are subsequently created in the target lexicon are annotated with their relation to entries from the original lexica. This represents a kind of 'log' of the merging process. One application of this is to allow the linguist to find justification for the existence of new lexical entries in terms of the original lexica.

5. Future Steps

Since people want to work on lexica not only via the web, but also on their notebooks, we will create a local version of LEXUS as well. LEXUS is written in Java and can be configured to use several database systems such as the Postgres database system or MySQL. The task is to create a suitable packaging that hides all complexity of the underlying application components to the user who will install the tool.

To increase its flexibility we will add some functionality that allows users to more easily integrate lexica that are represented in XML although there may be not an XML schema.

It is obvious that we will have to adapt the user interface several times and offer different shells around the core. Some researchers want to have simple user interface so that they can interact with their consultants directly and establish remote collaborations. In this respect we are planning to start a concrete project for building up a multimedia lexicon with many relation types between lexical entries and attributes. In such a project where we will sit together with community members we

will have to understand what simplicity in their eyes will mean.

The started merging work will continue to be able to offer more advanced merging capabilities. Given the increase of remote collaborations and the wish to merge and relate lexica from different origins we foresee that there is an increasing need in efficient merging functionality.

Another step that has already been started is to let LEXUS interact with ANNEX [9] which is the web-based annotation utilization tool built at the MPI. An excellent example for interaction functionality is offered by Shoebox. We have to see which kind of interactions will be useful for our users. First simple forms were implemented so that it is possible to select a string in an annotation and then invoke the appropriate lexicon entry and vice versa.

6. Conclusions

LEXUS can fulfill an excellent role for accessing lexica that are stored in an archive and that were delivered in various structures and formats. Its interoperability functionality allows users to operate in a cross-lexicon manner, although true ontology support has to be added. It is a flexible tool that allows users to create his own structure, his own choices of lexical attributes where he can chose between concepts already defined in registries or own definitions and his way of presenting the data on screen and on paper. We have the intention to add more advanced functionality so that LEXUS is not only a tool supporting the LMF standard, but also offers useful services for the researcher not yet provided by other tools.

7. References

[1] http://www.mpi.nl/DOBES
[2] http://www.sil.org/computing/shoebox
[3] http://www.ru.nl/celex
[4] http://www.tst.inl.nl/cgn.htm
[5] Wittenburg, W. Peters, S. Drude (2002): *Analysis of Lexical Structures from Field Linguistics and Language Engineering*. LREC 2002 Conference. Las Palma, Mai
[6] http://stp.ling.uu.se/~joerg/diplom/node5.html
[7] http://www.w3.org/2001/sw/BestPractices/WNET/IdeLenci.pdf
[8] www.tc37sc4.org
[9] atoll.inria.fr/RNIL/TC37SC4-docs/N089.pdf
[10] http://www.ub.es/gilcub/SIMPLE/simple.html
[11] http://www.ub.es/gilcub/SIMPLE/simple.html
[12] http://www.mpi.nl/lexus
[13] http://www.cs.vassar.edu/~ide/papers/LREC2004-DCR.
[14] http://www.loria.fr/projets/TMF/tmf.html
[15] http://nlp.stanford.edu/kirrkirr
[16] http://www.sil.org/computing/shoebox/
[17] http://childes.psy.cmu.edu/
[18] http://www.mpi.nl/annex