Creating multimedia dictionaries of endangered languages using LEXUS

Jacquelijn Ringersma[1] *and Marc Kemps-Snijders*[1]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
jacquelijn.ringersma@mpi.nl, marc.kemps-snijders@mpi.nl

Abstract

This paper reports on the development of a flexible web based lexicon tool, LEXUS. LEXUS is targeted at linguists involved in language documentation (of endangered languages). It allows the creation of lexica within the structure of the proposed ISO LMF standard and uses the proposed concept naming conventions from the ISO data categories, thus enabling interoperability, search and merging. LEXUS also offers the possibility to visualize language, since it provides functionalities to include audio, video and still images to the lexicon. With LEXUS it is possible to create semantic network knowledge bases, using typed relations. The LEXUS tool is free for use.

Index Terms: lexicon, web based application, endangered languages, language documentation.

1. Introduction

LEXUS [1] is web-based lexicon tool developed at the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands (MPI). It is targeted at linguists doing field research or working with language corpora. LEXUS is of primary interest for language documentation projects, since it offers the possibility to not just create a (digital) dictionary or thesaurus, but additionally it allows creating a multimedia encyclopedic lexicon.

The idea for LEXUS originates from the language documentation program (DOBES) funded by the Volkswagen Foundation [2]. Within the DOBES projects [3] broad collections of spoken text genres in various contexts and socio-cultural interactions are collected in the form of audio and video recordings in order to document languages that are potentially in danger of becoming extinct within a few years time. The DOBES pilot phase started with seven documentation teams, and first analysis showed that almost every researcher created lexica dependent on the language and linguistic theories in focus. This resulted in an interoperability problem due to a total of 10 different lexicon structures, variation in attribute and value naming (e.g. the concept 'noun' can be labeled as 'N', 'no' and 'noun'), and a diversity of formats and tools used for the creation of lexica (Toolbox, Access, Excel and Word, XML).

At the same time an analysis by ISO for the language engineering field brought forward the conclusion that there was a need for standardization of linguistic resources, which resulted in the creation of the ISO TC37/SC4 group on standardization in linguistic resource management [4]. The recommendations of this group led to: (1) the definition of a lexicon framework which is flexible enough to handle all possible lexicon structures and (2) the creation of a Data Category Registry containing standards on lexicon attributes and values. The use of such standards allows interoperability, for instance merging of lexica, cross lexica searches and semantic unification. LEXUS is an implementation of these recommendations.

LEXUS is not just another lexicon tool: LEXUS offers the possibility to visualize language through the import of multimedia. Four different types of multimedia fragments can be linked to the lexical entries: drawings, photos, videos and audio files. In addition typed relational linking between lexical entries is supported to include information such as examples found in other (multimedia) documents, structural dependencies, semantic references, etc. Such typed relations can amount to knowledge and semantic networks dependent on the users' intentions.

LEXUS also interacts with media stored on the Internet. For the DOBES projects data is stored in the digital archive for linguistic resources housed at the MPI [5]. The archive is accessible via the Internet, and is organized in a structured manner by describing and contextualizing the data with the IMDI metadata set [6]. Lexical entries in LEXUS can be linked to domains and resources in this archive.

LEXUS is a web based tool, with a stand alone version available for users without Internet access. Lexica and lexicon structures may be exported in XML format. LEXUS is free for anyone to use through www.mpi.nl/lexus.

2. The LEXUS pilot projects

For the initial developments of LEXUS we have been working in close collaboration with the users from the DOBES projects. For the first lexicon implementations we are further developing LEXUS in close cooperation with two documentation projects: (1) the DOBES project 'Towards a multimedia dictionary of the Marquesan and Tuamotuan languages of French Polynesia' and (2) the MPI project for the creation of the Yélî Dnye dictionary and semantic network. Both projects are a cooperation between linguists and technicians, however in order to create a user interface (UI) which is adjusted to the IT skills and knowledge levels of the potential end users, the first project also involves the Marquesan and Tuamotuan speech communities in the LEXUS development.

2.1. Marquesan and Tuamotuan dictionary

Marquesan and Tuamotuan are two languages spoken on the Marquesan islands of French Polynesia. From 2003 to 2005, a broad variety of spoken text genres in form of audio- and video-recordings of these languages was collected in the DOBES project 'Documentation of the Marquesan languages and culture in French Polynesia'. The recordings are stored in the MPI digital archive for linguistic resources, structured according to wide range of topics such as story-telling, song and dance, traditional food preparation, plant medicine, fishing techniques, aspects of the material culture and artifacts, traditional practices and the use of various trick

languages [7]. The documents have been transcribed and translated together with native speakers.

For the Marquesan language a trilingual general dictionary (Marquesan vernacular, French, English) with thematic glossaries of topics, such as food preparation and conservation, plant medicine, fishing and breadfruit varieties, has been created in Toolbox [8]. Since idioms provide a deep insight into a culture and constitute part of the linguistic competence of speakers [9] the dictionary and glossaries are further complemented by a collection of idioms and collocations which are in danger of disappearing as the Marquesan vernaculars are undergoing rapid linguistic change in the younger generations.

Within the framework of the DOBES project 'Towards a multimedia dictionary of the Marquesan and Tuamotuan languages of French Polynesia' we are integrating the lexicon in LEXUS, including the multimedia elements into the structural frame of a conventional dictionary. By going beyond traditional practices and theoretical considerations in lexicography, these elements can also represent the meaning of words in a new way. While multimedia enrichments document the meaning of words more completely in its indigenous context, the creation of typed relations of various sorts will place words in their semantic contexts with other words, with examples in annotations.

2.2. Yélî Dnye lexicon and classification of the natural world

Yélî Dnye is a Papuan language spoken on Rossel Island, Louisiade Archipelago, Papua New Guinea.

Language documentation of Yélî Dnye has been taking place since 1995 and part of this documentation takes place within the project Pioneers of Island Melanesia [10]. A Toolbox dictionary has been created containing over 6000 lexical entries, a large part of them nouns representing objects and entities in the natural world. Although no systematic ethnobiology has been done on Rossel, and therefore the biological taxa have not been established we are using LEXUS to tentatively create a semantic network visualizing a classification of the natural world [11]. After this initial creation, the use of LEXUS collaborative workspaces will allow the speech community to improve and extend this classification.

Lexical entries will be completed and enriched with multimedia elements and links to the digital archive domain and resources, showing the objects and entities within their natural context.

3. LEXUS functionalities

3.1. New lexica

The LEXUS lexicon structure is based on the ISO TC37/SC4 Lexical Markup Framework (LMF). LMF is a generic model allowing users to define almost any type of structure for their lexicon, from simple word lists to complex multi-lingual lexica. In LMF the default lexicon structure consists of two components, one for the general information on the lexicon (lexiconInformation) and one for the structure of the lexical entries. Lexical entries consist of a Form and a Sense component. Components are containers of data categories: the actual lexical attributes, which can be valued for each individual lexical entry. Users may define the structure of the default components according to the linguistic theories or requirements of the language under research. Figure 1 gives an example of such user defined structure for the Form component. In the example Form consist of two sub-components: inflectedForm and lemmatizedForm. Both of these sub-components contain three data categories which the user defined to specify the characterics of the two Forms. LEXUS also allows the user to use well accepted data category registries (ISO, Toolbox MDF) for the naming of the data categories. The example in Figure 1 shows the use of the ISO concept naming for the Form data categories.

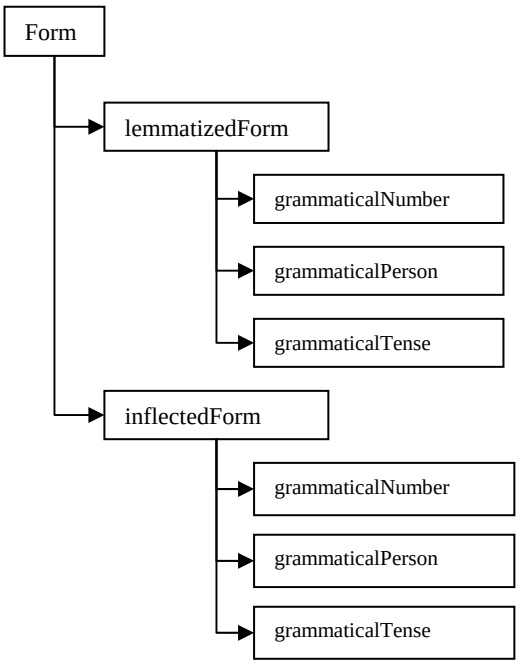

Figure 1: *Schematic diagram of a user defined LMF Form component*

LEXUS allows the creation of lexica from scratch, however since LEXUS is primarily meant for field linguists and since the Toolbox [8] is a widely used data management and analysis tool for field linguists, LEXUS also supports the import of lexica created in Toolbox.

Both the pilot projects have lexica created in Toolbox. Lexicon structures are built in the Toolbox database type (MDF), which includes linguistic field markers for lexeme, part of speech, definition, vernacular gloss etc. The head marker is the lexeme marker and the other markers are structured under this head marker. The structure includes the internal textual organization and the information structure of the lexical entries. Besides this linguistic information, non-linguistic markers are included to provide encyclopedic information, like e.g. the scientific name of objects in the natural environment. LEXUS supports the import of the Toolbox structure (the so-called type files): the markers are imported as data categories and the structure is implemented by grouping the different data categories under a data component. Users may specify the location of a Toolbox marker in the LMF model: a marker can be placed under lexicalEntry, Form or Sense. LEXUS maintains the structures underneath these markers. The Toolbox lexicon data file is stored as plain text 'database' file. LEXUS imports this file and creates the LEXUS lexical entry values for the data categories.

LEXUS also supports the import of XML formatted lexica and lexica created with the CLAN software used in

acquisition linguistics. Both import types allow the creation of the lexicon structure under the LMF core model, similar to the Toolbox import.

3.2. Viewing the lexicon

The LEXUS main window, consists of an alphabetic (or otherwise ordered) wordlist. When creating a new lexicon, by default the lexeme data category represents the lexical entry in the wordlist. The creator of the lexicon, however, can add other attributes (e.g. lexical elements and markup) for this representation. For the Marquesan lexicon we have selected the 'lexeme', 'Part of Speech', 'Definition (E)' and 'Definition (n)' attributes to represent a lexical entry in the word list. This selection facilitates quick browsing of the lexicon. Details of the lexical entry can be viewed by selecting an entry from the word list. This opens a new window with a full view of the lexical entry. Also for this view the creator is free to select the attributes to represent the lexical entry (see Figure 2).

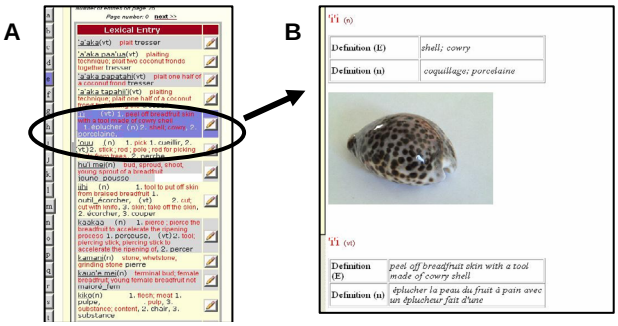

Figure 2: *Ordered word list (A) showing the lexeme and descritpion for the lexical entries and the full view (B) of the lexical entry 'i'i*

3.3 Multimedia

To easily identify multimedia content in the lexicon, additional data categories for the different types of media have been included in the structure of the lexicon. For both pilot projects we created one data component containing four data categories: audio, video and photo and drawing. When selecting one (or more) of these data categories in the full view of the lexical entry, the lexical entry will be displayed together with the selected media object (see Figure 2).

For all headwords denoting an object in the natural world we have foreseen a value for the photo data category. In the Marquesan lexicon, some lexical entries will also be represented with a drawing. For example the breadfruit variety *mei* as well as a number of other plants in the Marquesan culture can be used in various ways. It is always a specific part of the plant which is used for a specific purpose. Photos cannot always show the required detail to visualize certain characteristics of a plant, which is why we have chosen to use drawings (see Figure 3).

Besides the visual media we linked sound files to the lexical entries. In the full view of the lexical entry it will be possible to play the sound file, giving the user a possible pronunciation of the headword or a sample sentence.

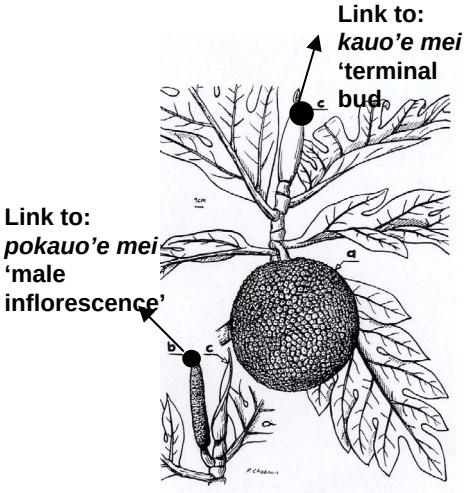

Figure 3: *Image of the Marquesan lexical entry 'mei' with links to details of specific parts in order to visualize functional characteristics.*

3.4 Linking to digital (language) archives

The data of both pilot projects are stored in the MPI archive for linguistic resources [5]. The archive stores the video and audio files with their ELAN [12] annotations (transcription, translation, comments). Each file in the archive is identifiable with a persistent identifier (URI/handle). LEXUS uses this handle to link the stored data as values for the audio and video data categories. Not only is it possible to link the whole file, linking selections within the video files is also an option. The advantage of the archive linking is that the LEXUS lexicon file remains relatively small (easy manageable) since a large amount of the data is stored outside LEXUS. The disadvantage is that the multimedia extensions in the lexicon are only readable for those people who have access to the Internet and who have been granted access rights to the resource files in the archive (see Figure 4).

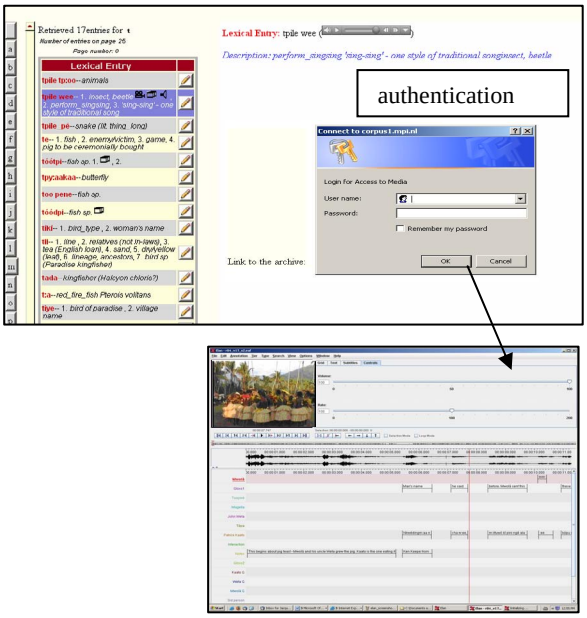

Figure 4: *LEXUS link to archived media and ELAN annotation file (Yélî Dnye lexicon).*

3.5 Relational networks

Semantically annotated words, which are part of relational networks, offer an intuitive entry to the lexicon for users other than linguists. Words appear as part of the conceptual world representing part of the indigenous knowledge.

LEXUS allows the user to create these semantic networks (see Figure 5) but also to navigate through the lexicon using these networks, leaving the viewer of the lexicon free to find his way through the lexicon following the path of his personal interest only. Relational links in LEXUS are typed: they have attributes defining define the type and directionality of the relation. The creator of a lexicon is free to define his own relation types, but LEXUS also offers some default paradigmatic relations: synonymy, antonymy, hyponymy and meronymy.

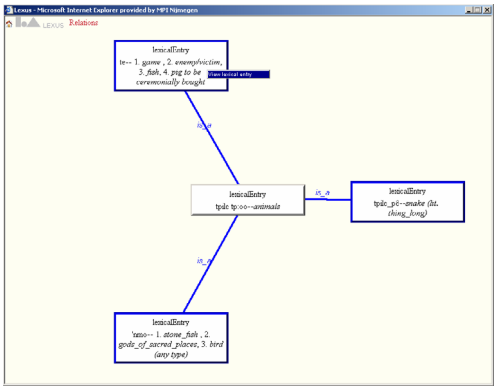

Figure 5: *(Part of) semantic network in Yélî Dnye lexicon*

4. Future developments

4.1 Collaborative workspaces

In LEXUS, lexica are created in secure workspaces, which require user authentication. The user may publish the lexicon to a central storage and define access rights to users who may retrieve the lexicon into their own workspaces.

We are currently working on the realization of a workspace concept allowing collaborative lexicon creation. Collaborative lexicon creation can be realized when more than one person is able to work on the same lexicon in different workspaces at the same time. This allows different researchers as well as members of the speech community to contribute in the construction of a rich resource for language documentation and revitalization. Collaborative lexicon creation in multiple and simultaneous workspaces requires flexibility as well as mechanisms to merge and consolidate the enriched lexicon versions. Since the people involved work at different locations, this collaboration has to be built on a virtual space that is accessible by all, which is the Web.

4.2 User Interface

The final LEXUS challenge is the creation of a user interface, adjusted to the knowledge and (IT) skills of the different speech community. Such a UI is an essential prerequisite for participation of this community in the creation of lexica. We identify an interesting software engineering phenomenon: normally a user interface design is focusing on bridging the built-in tool functionalities and cognitive ergonomic principles. For an adjusted interface there is no such convergence towards an 'optimal solution', since the concept of 'adjusted' depends on the selection of functionality and on cultural preferences The creation of such an adjusted interface can be achieved only in close collaboration between the researchers, developers and speech community members.

5. Conclusion

LEXUS is a flexible lexicon tool under development at the Max Planck Institute for Psycholinguistics. LEXUS functionalities currently allow the creation of digital, multi-lingual, multimedia dictionaries. These functionalities include the creation of LMF structured lexica, import of Toolbox and XML lexica, the integration of multimedia fragments and the creation of semantic networks representing indigenous knowledge bases.

The next step in the development of LEXUS consists of the implementation of collaborative workspaces, allowing multiple users to create a lexicon simultaneously. This functionality will require mechanisms to merge the different lexicon versions. Also, we have now arrived at a stage in the development phase, in which we will concentrate on the improvement of the LEXUS user interface. For this activity we have foreseen an important participation of the speech community. Representatives of the Marquesan and Tuamotuan speech communities will visit the Max Planck Institute in the summer of 2007, in order to facilitate the developers of the tool to adjust the user interface to the knowledge level and IT skills of the future users of the lexicon.

We plan to deliver LEXUS to the speech community mid 2008. The current LEXUS version (0.93) is available from: http://www.mpi.nl/lexus.

6. References

[1] LEXUS, "A web based lexicon tool" url: http://www.mpi.nl/lexus 2006.

[2] Volkswagen foundation. url: http://www.volkswagenstiftung.de/

[3] DOBES. Documentation of endangered languages. url: http://www.mpi.nl/dobes/. 2006

[4] ISO/TC 37/SC 4 Committee. Lexical Resource Markup Framework (LMF). http://iso.nocrew.org 2003

[5] MPI, 2007. Digital Archive for Linguistic Resources. url: http://corpus1.mpi.nl/ds/imdi_browser/

[6] Wittenburg P., W. Peters, and D. Broeder, Metadata proposals for corpora and lexica. Proceedings LREC 2002, Las Palmas pp.1055 – 1059

[7] Cablitz G. Towards a multimedia dictionary of the Marquesan and Tuamotuan languages of French Polynesia. Project Application to the Volkswagen foundation. Germany. Personal Archive. 2006.

[8] Toolbox, SIL. The Linguist's Shoebox. url: http://www.sil.org/computing/catalog/

[9] Pawley A.K., A language which defines description by ordinary means. In: Foley A.F. (ed.) The role of Theory in Language Description, 87 - 129. Mouton de Gruyter. Berlin. 1993

[10] Pioneers of Islands Melanesia http://www.eastpapuan.ling.su.se/

[11] Levinson, S.C. A Grammar of Yélî Dnye, the Papuan Language of Rossel Island. In preparation.

[12] ELAN, Extended Linguistic Annotator. url: http://www.mpi.nl/tools/elan.html 2006