Visualizing endangered indigenous languages of French Polynesia with LEXUS

Gaby Cablitz[1], Jacquelijn Ringersma[2] and Marc Kemps-Snijders[2]
[1] Seminar für Allgemeine und Vergleichende Sprachwissenschaft der CAU, Kiel, Germany
[2]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
{gcablitz@linguistik.uni-kiel.de, jacquelijn.ringersma@mpi.nl, marc.kemps-snijders@mpi.nl}

Abstract

This paper reports on the first results of the DOBES project 'Towards a multimedia dictionary of the Marquesan and Tuamotuan languages of French Polynesia'. Within the framework of this project we are building a digital multimedia encyclopedic lexicon of the endangered Marquesan and Tuamotuan languages using a new tool, LEXUS. LEXUS is a web-based lexicon tool, targeted at linguists involved in language documentation. LEXUS offers the possibility to visualize language. It provides functionalities to include audio, video and still images to the lexical entries of the dictionary, as well as relational linking for the creation of a semantic network knowledge base. Further activities aim at the development of (1) an improved user interface in close cooperation with the speech community and (2) a collaborative workspace functionality which will allow the speech community to actively participate in the creation of lexica.

Keywords --- **Language documentation, multimedia lexicon, endangered languages**

1. Introduction

Language documentation is a new field in linguistics dealing with methods, tools and theoretical backgrounds for compiling a representative and long lasting, multipurpose record of natural languages [1]. Currently there are approximately 6000 languages in use worldwide. However, by the end of the 21st century, only one half of these languages will continue to exist [2]. Documentation work can help to maintain, consolidate or revitalize endangered languages, safeguard the full range of uses of a given language and inform future generations about the language diversity and the cultural treasures of mankind [3]. Documentation should be carried out in close cooperation with the speech community and reflect the particular characteristics of the respective culture.

The role of data in language documentation differs from the way data is treated in language description in that speech communities and future researchers can fully reconstruct the social and cultural contexts with the help of the documented material [1]. Projects within the Volkswagen foundation initiative for the Documentation of Endangered Languages (DOBES) build up language archives of primary data: audio recordings present the languages as they are spoken with all their segmental and suprasegmental richness; video recordings give additional information about the socio-cultural environment and non-verbal communication practices. The secondary material, derived from these recordings, are transcriptions, translations, morphosyntactic and lexical analysis and possibly general comments on content and linguistic phenomena. Both the recordings (and in addition photographic material) and the annotations are stored in digital form.

A number of digital archives exist [4], offering storage space for data from communities and researchers, as well as access to (future generations of) speech communities, researchers and the general public. For the DOBES projects data is stored in the archive for linguistic resources housed at the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands [5]. The digital 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]. The archive takes care of the long-term persistency of the digital material, which is not evident given the limited life-time of our state-of-the-art storage media. For specific media data highly specialized viewers have been developed to support direct access to the media content. An example of such a viewer is ANNEX [7] for visualizing annotated media files. ANNEX provides different views of the annotations and streams media through the QuickTime plug-in in the browser.

In 2003 the project 'Documentation of the Marquesan languages and culture in French Polynesia' started to document the endangered Marquesan languages in various different contexts and socio-cultural interactions.

The major aim of this project was to collect a broad variety of spoken text genres in form of audio- and video-recordings in order to document the language of various endangered cultural practices. The Marquesan archive covers a 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 (fabrication of bark cloth, traditional tool-making), traditional practices (e.g. life cycle) and the use of various trick languages [3]. The documents have been transcribed and translated together with native speakers. Translations of the transcripts exist in French as well as English. Comments and notes on cultural phenomena and linguistic structures are annotated in English.

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. Since idioms provide a deep insight into a culture and constitute part of the linguistic competence of speakers [8] 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 building a digital multimedia lexicon of the endangered Marquesan and Tuamotuan languages with the help of the new web-based lexicon tool, LEXUS which allows the integration of multimedia elements into the structural frame of a conventional dictionary [9]. The multimedia lexica are created on the basis of the databases developed in the previous project. By going beyond traditional practices and theoretical considerations in lexicography, they can also represent the meaning of words in a new way. For example, video clips of motion verbals – designed and acted out by native speakers – shall give deeper insights into the speakers' knowledge systems and representation of meanings. Furthermore, the writing of vernacular definitions by members of the speech communities will be an important step towards language maintenance and revitalization as well as indigenous knowledge representation of word meaning. 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 etc. LEXUS is a web-based tool targeted at field linguists. With LEXUS users are able to construct flexible lexicon structures and include multimedia elements to the lexical entries and it also allows the creation of semantic networks, which makes LEXUS the ultimate tool for the creation of encyclopedic dictionaries.

The project is a cooperation between linguists, technicians and the speech communities in which the LEXUS user interface is being developed in such a way that it is adjusted to the wishes and knowledge of the speech community. Visualization functionalities are being improved in order to be able to present the data in an ethnographic way adopted to the needs of the speech community. The creation of semantic networks is being facilitated so that both linguists and members of the speech community are able to draw functional relations between the lexical entries and cultural/indigenous concepts in the lexicon. This paper reports on the background and philosophy of the project, gives a description of the lexicon tool and presents the first results of the project.

2. LEXUS

LEXUS [9] is a web-based lexicon tool developed by the Max Planck Institute for Psycholinguistics. 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 to create a multimedia encyclopedic lexicon. To achieve this LEXUS supports flexible lexicon structures, Shoebox [10] compatibility, a flexible scheme of linking with multimedia documents and the possibility to include relational links between entries or attributes of entries. LEXUS is able to interact with ANNEX for the visualization of annotated media files. The utilization of electronic multimedia archives offers new ways of making visualized dictionaries. Furthermore LEXUS is a new component in archiving, since it has possibilities of linking and interlinking linguistic as well as cultural concepts, thus creating a dense network of indigenous knowledge and concepts which has not been achieved in conventional electronic language archiving so far.

Being based on the Lexical Markup Framework (LMF) LEXUS is compliant with the proposed ISO standard for linguistic resources (ISO TC37/SC4). 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 actual lexical entries.

Lexical entries consist of a Form and a Sense component. Users may further construct the structure of the lexicon or lexical entry according to the specific structure of the documented language or the linguistic theory used.

LEXUS allows the creation of lexica from scratch as well as the import of lexica created in Shoebox (or the newer version of the tool: Toolbox). In addition lexica formatted in XML can be imported in LEXUS, provided that the XML schema is available.

Four different types of multimedia fragments can be linked to the lexical entries: drawings, photos, videos and audio files. Links can also be created to archived media files and annotations. Within the MPI archive [5], this means that LEXUS allows interaction with the visualization tool ANNEX [7]. In addition, typed relations are supported to include information such as examples found in other (multimedia) documents, structural dependencies as they occur in morpho-syntax, semantic references, etc. Such typed relations can amount to networks or taxonomies dependent on the users' intentions.

LEXUS allows web-based access and collaboration of several researchers as well as people from the speech communities of (endangered) languages. For this purpose LEXUS offers to create workspace copies for different people and a merging of different versions. Thus, researchers may evaluate contributions and modifications and create master copies at regular intervals.

The tool opens the possibility of actively involving people from the language communities in adding rich semantic/encyclopedic information which linguists alone will not be able to achieve.

3. Design of a multimedia dictionary

3.1 Creating the basics: importing the Shoebox lexicon into LEXUS

Within the ongoing 'Documentation of the Marquesan languages and culture in French Polynesia' project a trilingual lexicon has been created in Shoebox [10]. The structure of this lexicon is built in the MDF 4.0 Shoebox database type, and includes linguistic field markers for lexeme, part of speech, definition (E), definition vernacular, gloss (E), gloss (national language) etc.

The head marker is the lexeme marker (\lx) and the other markers are all 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 Shoebox structure: the markers are imported as data categories and the structure is implemented by grouping the different data categories under a data component. The values 'attached' to the Shoebox markers are attributed to the data categories in LEXUS.

The LEXUS main window, consists of an alphabetic (or otherwise ordered) wordlist (see Figure 1). When importing a Shoebox lexicon, by default the head marker will represent the lexical entry in the wordlist. The creator of the lexicon, however, can add other attributes (e.g. lexical elements) 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 a quick scan of, and search in, 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. For our project we initially selected the same as we selected for the word list, but in a different format. In a later stage we also added an image of the object to this view (see Figure 1).

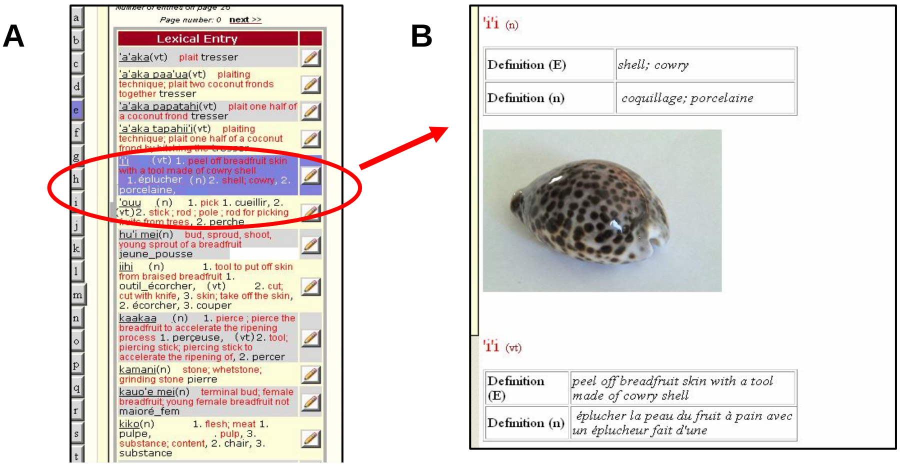
3.2. Including media into the lexicon

To easily identify multimedia content in the lexicon, LEXUS requires that additional data categories for the different types of media are included in the structure of the lexicon.

For the Marquesan lexicon three types of media were collected and will be linked to the lexical entries: audio recording, video recording and still images. In LEXUS 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 1).

For all headwords denoting an object in the natural world we have foreseen, minimally, a value for the photo data category. 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 2).



Number of entries on page 25

Page number: 0 next >>

Lexical Entry	
'a'aka(vt) plait tresser	
'a'aka paa'ua(vt) plaiting technique; plait two coconut fronds together tresser	
'a'aka papatahi(vt) plait one half of a coconut frond tresser	
'a'aka tapahii'i(vt) plaiting technique; plait one half of a coconut frond by hitching the tresser	
'i'i (vt) 1. peel off breadfruit skin with a tool made of cowry shell 1. éplucher (n) 2. shell; cowry, 2. porcelaine,	
'ouu (n) 1. pick 1. cueillir, 2. (vt) 2. stick ; rod ; pole ; rod for picking fruits from trees, 2. perche	
hu'i mei(n) bud, sproud, shoot, young sprout of a breadfruit jeune_pousse	
iihi (n) 1. tool to put off skin from braised breadfruit 1. outil_écorcher, (vt) 2. cut; cut with knife, 3. skin; take off the skin, 2. écorcher, 3. couper	
kaakaa (n) 1. pierce ; pierce the breadfruit to accelerate the ripening process 1. perçeuse, (vt) 2. tool; piercing stick; piercing stick to accelerate the ripening of, 2. percer	
kamani(n) stone; whetstone; grinding stone pierre	
kauo'e mei(n) terminal bud; female breadfruit; young female breadfruit not maioré_fem	
kiko(n) 1. flesh; meat 1. pulpe, . pulp, 3. substance; content, 2. chair, 3. substance	

'i'i (n)

Definition (E)	shell; cowry
Definition (n)	coquillage; porcelaine

'i'i (vt)

Definition (E)	peel off breadfruit skin with a tool made of cowry shell
Definition (n)	éplucher la peau du fruit à pain avec un épluchuer fait d'une

Figure 1: Ordered word list (A) showing the lexeme and descritpion for the lexical entries and the full view (B) of the lexical entry *'i'i*

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.

For headwords denoting complex cultural phenomena, practices and products there are several ways to include video files. One possibility is to link the lexical entry directly to a video file, appearing in the lexical entry view together with a transcription, translation and comments (for which we use the ELAN multimedia annotator [11]). However video files can be too large to load into PC systems or might be too lengthy for the user to watch the whole video.

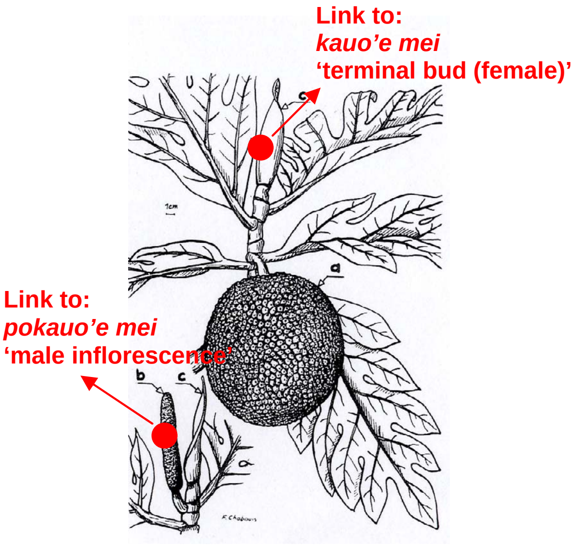


Figure 2: Image of *mei* with links to details of specific parts in order to visualize functional characteristics.

Therefore we opted to: (1) divide the video into (timed) sequences of the crucial elements or steps in the procedure or the practice, (2) link the headword to a photo gallery depicting the element or step and (3) link each photo to the respective video sequence.

We can illustrate this second approach with the headword *maa*. *Maa* is the Marquesan name for fermented breadfruit. The process of breadfruit fermentation consists of six steps: *hakapa'a te mei* **(**ripen the breadfruit), *'a'aka te po'a* (plait coconut leaves), *kanea te tapo'o maa* **(**make a container for *maa*), *kanea te ve'eve'e mei* **(**making tool for peeling breadfruit), *veve'e te mei* (peeling semi-ripe breadfruits) and *tata te mei* **(**remove ripe breadfruit from its heart and put into *maa*-container). The lexical entry *maa* will get six values for the photo data category, one photo for each step. In the lexical entry view *maa* all six photos (plus the denotations) will be shown. From these photos links to the videos (together with a transcription, translation and comments) will be made, so that the user can watch the whole *maa* preparation process in a sequence of short videos. Also each image is linked to the respective lexical entries, so that the user can jump from the *maa* entry to the lexical entries denoting part of the process.

The third type of lexical entries consist of proper names of locations and historical and contemporary persons. These occur primarily in narratives. Locations or sites which play important roles in narratives, and additional explanations on protagonists and historical personalities occurring in narratives have been filmed during the documentation project. This video material will be linked with the lexical entries.

3.3 Linking to digital (language) archives

The data of the ongoing project 'Documentation of the Marquesan languages and culture in French Polynesia' are stored in the DOBES [3] domain of the MPI archive for linguistic resources [5]. The archive stores the video and audio files with their ELAN [11] or Shoebox 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 (provided that the required time coding is known to the creator). 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 3).

4. Further developments

4.1 Visualizing semantic networks

LEXUS allows the creation of semantic networks. Semantic networks can be seen as forms of knowledge representation in which the relation between the concepts is visualized in a directed graph. Semantically annotated words, which are part of relational networks, offer a natural entry point for users other than linguists, since words appear as part of the conceptual world representing part of the indigenous knowledge rather than just a word in a lexicon. LEXUS allows the user to create these semantic networks 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.

The relational links in LEXUS have attributes which 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 [12]. A synonym set includes word concepts which all have the same semantic properties but different forms (war – armed conflict), the relation type synonym is bi-directional. Antonyms are words in a pair with opposite meanings (white – black), also this type is bi-directional. Hyponymy is a 'is a kind of' relation (horse – animal). A meronym denotes a constituent part of something (arm – body).

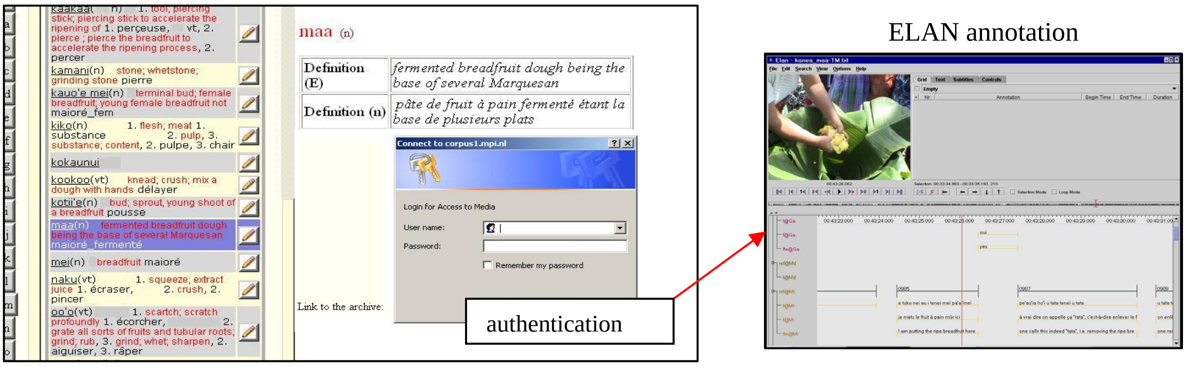
In the Marquesan multimedia lexicon we envisage the creation of a semantic network which represents the natural world of the indigenous speech community.

The intention is to involve the speech community as much as possible in the creation in order to obtain an ethno-biological or emic representation of the natural world, meaningful to the future users of the lexicon.

4.2 Involvement of the speech community

4.2.1 User Interface

Due to the involvement of members of the speech community, a user interface adjusted to the knowledge and (IT) skills of this community is required. The creation of such an adjusted interface can be achieved only in close collaboration between the researchers, developers and speech community members. Here we identify an interesting software engineering phenomenon. The standard user interface design is focusing on bridging the built-in functionality of a tool and general ergonomic principles so that one could speak about the existence of an 'optimal solution'.



kaakaa(n) 1. tool; piercing stick; piercing stick to accelerate the ripening of 1. perçeuse, vt, 2. pierce ; pierce the breadfruit to accelerate the ripening process, 2. percer

kamani(n) stone; whetstone; grinding stone pierre

kauo'e mei(n) terminal bud; female breadfruit; young female breadfruit not maioré_fem

kiko(n) 1. flesh; meat 1. substance 2. pulp, 3. substance; content, 2. pulpe, 3. chair

kokaunui

kookoo(vt) knead; crush; mix a dough with hands délayer

kotii'e(n) bud; sprout, young shoot of a breadfruit pousse

maa(n) fermented breadfruit dough being the base of several Marquesan maioré_fermenté

mei(n) breadfruit maioré

naku(vt) 1. squeeze; extract juice 1. écraser, 2. crush, 2. pincer

oo'o(vt) 1. scartch; scratch profoundly 1. écorcher, 2. grate all sorts of fruits and tubular roots; grind; rub, 3. grind; whet; sharpen, 2. aiguiser, 3. râper

maa (n)

Definition (E)	*fermented breadfruit dough being the base of several Marquesan*
Definition (n)	*pâte de fruit à pain fermenté étant la base de plusieurs plats*

Link to the archive:

Connect to corpus1.mpi.nl

Login for Access to Media

User name:

Password:

□ Remember my password

Figure 3: Lexical entry view for *maa*, with link to the digital archive. The ELAN annotation (including the video) opens after authentication.

For an adjusted interface there is no such convergence towards an optimal solution, since the concept of 'adjusted' is very much dependent on a personal selection of functionality and on personal preferences. An 'adjusted' interface, therefore, is a unique implementation which takes account of the wishes of a specific group of persons or even individuals as long as there are no simple standards established by broad training and education.

4.2.2 Collaborative workspaces

In LEXUS, lexica are created in, so called, workspaces. To enter into a workspace the user needs to authenticate with a username and password. After the creator of the lexicon is satisfied with his or her product the lexicon can be published to a central storage. When publishing the creator needs to set the access rights to the lexicon, meaning that he needs to specify the user or group of users which can import the lexicon to their workspaces and make modifications. Within the framework of the project we want to realize the concept of collaborative lexicon creation. Collaborative lexicon creation can be realized when more than one person can work on the same lexicon in different workspaces at the same time, allowing members of the speech community to freely contribute in the building up of a rich resource for language maintenance measures and further research on the language and culture. For the researcher this work by indigenous writers will expand their research sources and will give them new tasks in guiding, moderating and unifying the commentaries and extensions.

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.

5. Conclusion

LEXUS is a flexible lexicon tool under development at the Max Planck Institute for Psycholinguistics. Within the framework of the project ‘Towards a multimedia dictionary of the Marquesan and Tuamotuan languages of French Polynesia’ we have developed several functionalities in LEXUS which allow the creation of a digital, multi-lingual, multimedia dictionary. These functionalities include the import of Shoebox lexica, the integration of audio and video fragments and the linking of images to lexical entries. Furthermore, with LEXUS the user is able to create semantic networks representing indigenous knowledge bases in which relations between objects and entities are visualized in directed graphs.

Currently, we have 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 speech community 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.

The next and last 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 flexibility of the tool, as well as mechanisms to merge the different lexicon versions.

We plan to deliver LEXUS, with full functionalities and user interface to the speech community mid 2008. The current LEXUS version (0.93) is available from: http://www.mpi.nl/mpi/lexus.

References

[1] Gippert J., N.P. Himmelmann and U. Mosel (eds.), 2006. Essentials of language documentation. Mouton de Gruyter. Berlin.

[2] UNESCO, 2005. Language preservation. http://www.unesco.org

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

[4] DELAMAN, 2003. Digital Endangered Languages and Musics Archives Network. http://www.delaman.org/

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

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

[7] ANNEX, 2006. Annotation Exploration tool in the MPI web-based framework for archive exploration and enrichment. http://www.mpi.nl/annex/

[8] Pawley A.K., 1993. 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.

[9] LEXUS, 2006. Lexus, a web-based lexicon tool. http://www.mpi.nl/lexus/

[10] SIL, 2006. Shoebox, The Linguist’s Shoebox. http://www.sil.org/computing/catalog/

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

[12] Murphy, M.L., 2003. Semantic relations and the lexicon. Antonymy, Synonymy and other paradigms. Cambridge University Press.