E-MELD Workshop 2006

Digital Language Documentation
Tools and Standards: The State of the Art

Working Group 3: Lexicon creation

Members: Steve Abney, Dunstan Brown, Osten Dahl,
*Sebastian Drude, Susanna Imrie, Marc Kemps-Snijder,
Christopher Manning, *Mike Maxwell, Vivian Ngai
(* = co-chairs)

1   General Comments

The good news is, there are many lexicon applications and standards for field linguists, with more coming. The bad news is, there are many lexicon applications and standards for field linguists, with more coming.

The members of the working group agreed that there is a wide variety of applications currently available for doing lexicography in field situations, including programs for building lexicons from data (such as corpora). The obvious advantage of this is that there is a wide range of choices to fit a wide range of uses (although all agreed that no single program was ideal). The obvious disadvantage of this is that it becomes difficult to choose the "right" program for a given project. A less obvious disadvantage is that it is difficult for a new (and possibly better) lexicon tool to break into an already crowded field.

While less numerous, there are several standards (including *de facto* standards) for lexicon design. Ideally, there would be one, with some parts of that one standard being required, and other parts being optional. But since there are several standards, no one standard can claim to be the one true standard. As a result, portability of lexicons between projects, and the ability to compare lexicons from different projects (e.g. when doing historical comparative linguistics) is limited.

These points will come up in later sections.

2   The 'ecology' of lexical data in language documentation

Lexicography in the context of language documentation and language research in general involves much more than 'lexicon creation' (the title of our working group).

Lexical data is a particular data type which is somewhat intermediate between language documentation (with focus on collection of data such as annotated texts) and language description (focus on data analysis). This is because it simultaneously consists of a collection aiming at covering a domain (lexical units), but at the same time it contains 'derived' information that requires linguistic analysis.

Questions to be covered in connection with lexical data are:

- How do these data inter-operate with other data types (in particular, with annotated texts or with

grammatical descriptions)?

- Which kind of information is to be included in a database of lexical data (grammatical levels from phonetics to pragmatics, paradigms etc...), and how is it structured?

- How can specific pieces of lexical data be accessed and retrieved (searched, ordered, etc.)?

- In which formats is lexical data archived, disseminated, and displayed?

- Who will use lexical data, and how?

An evaluation of existing tools, and the design (or wish-lists for such) should be specific as to which aspects of creating and handling lexical data they address. In order to organize this wide field, one can use the 'ecology'-metaphor introduced by Jeff Good and distinguish different players and related data types and relations among them; in particular, activities; some of these can be ordered chronologically in steps. As a first approach, we offer the following graphic, which should be expanded; especially the different kinds of relations (here simply listed by numbers) could be distinguished and classified into different basic types.

C   C   A
TXT
4
8
1   9
2   LDB
3
3a   6
1a   LDB   LD output
7   10
5
St + Th   12   12
(L)DB   11
U
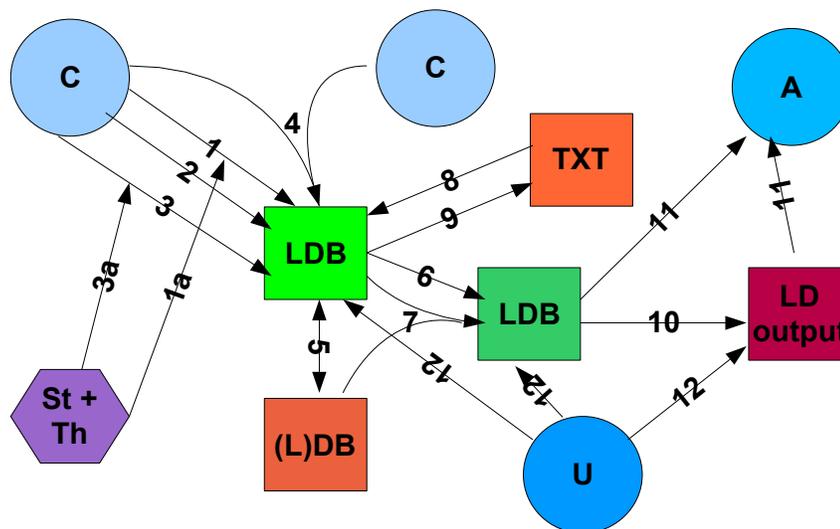
Explanations and comments:

- Circles: persons or institutions (need not to be different one from another)

  C: Creator, **U**: User, **A**: Archive(r)

- Squares: Data (in different formats, including hardcopies etc, usually digital files)

  LDB: Lexical DataBase, **(L)DB**: DataBase, of lexical or other data, **TXT**: Text database, **LD output**: Lexical data in dissemination format

- Hexagon: abstract entities, here: **St+Th**: Standards and Theories

In the center, there is the lexical database in light green color (and a second instance of it, in darker green, see below on (6) and (7); most relations may hold to both instances but are presented here only with respect one of the two).

1) The lexical database is created by a lexicographer (professional linguist or not, with different implications as to design of interface and sophistication etc.). The first step is the **DESIGN** of the database, this is, of the ordering and internal structure of the entries (macro- and micro-structure). A **lexicographic entry** is a data set about one **lexical unit**, which usually is conceived as a combination of a formal entity of the object language and one (or more) of its senses. But this conception and the details can vary among lexicographers of different convictions (for instance, what is the formal entity: a root, a stem, a characteristic word form, or an entire paradigm etc.) , which is why:

1a) linguistic theories and (technical and other) standards have a strong explicit or implicit **IMPACT** on the basic activities of lexical database creation. Tools vary as to the inter-theoretical applicability, that is, the flexibility to be used with and adjusted to the specific approach used by the lexicographers.

2) The main activity related to lexical databases is the **ADDING** of new entries (possibly with templates, or a fixed structure, or without any prescribed configuration...), and

3) adding and **EDITING** the different pieces of information (data categories, such as part of speech, definition, semantic relations, examples, etc.) related to the lexical unit treated. This, again, is:

3a) **BASED ON** the linguistic theories and standards underlying the lexicographic treatment.

4) In many cases, work on lexical databases is done or should be done **COOPERATIVELY** by several people (e.g., simultaneously by distributed editing, or by repeatedly merging different versions). Of course, tools vary in their ability to provide facilities for cooperative work.

5) Often lexicographic work is based on different databases, e.g. a second database for word forms (for instance a thesaurus or frequency lists extracted from text collections), or databases for grammatical classes and labels, or for registering semantic relations etc. Again, tools vary in their capacity to provide **INTEROPERABILITY** among different databases.

6) Lexical databases can be **TRANSFORM**ed or converted into other, derived databases (e.g. an index ordered by expressions in a target language, or a database in another technical format, such as Standard Format to XML-based format conversion).

7) Another frequent technical operation on lexical databases is the **MERGING** of two or more different databases into a new one (e.g., partial databases covering different semantic domains, or merging one database for affixes with another for free forms).

8) A special and often used type of inter-operation between different databases (see 5) is interaction with text databases. Most significantly, texts can be used for **EXTRACTING** (discovering) new lexical units and for providing natural **EXAMPLES**.

9) On the other hand, lexical databases are often used for **ENRICHING** annotation of texts, via consultation and especially for morphological parsing and semi-automatic glossing (different sub-processes of what is often called **INTERLINEARIZATION**).

10) When a certain stage of lexicon creation and enrichment is reached, the lexical database will be stored in some definite format, possibly different from the format used for editing and enriching or inter-operation with other databases, for instance for presentation or dissemination. **OUTPUT FORMATS** include a print-out (hard-copy) and electronic versions (on-line or standalone). Lexicographical tools again vary with respect to their capabilities of creating these.

11) Possibly the working format, the output format or even still another different format is used for **ARCHIVING** the lexical database, e.g. as part of a broad language documentation.

12) Finally, the destination of lexical databases is their **USE**, for different purposes such as looking up lexical entries for (active or passive) translation, language learning, linguistic research, support in cultural and historical studies etc. This use depends on the work-flow and organization of a research project as well on the tools applied which version of the lexical database a user (who need not be different from the creator(s) or archiver(s) of the databases in the first place) will be accessed by the user for retrieving lexical information.

In the time alloted to our working group, we could not completely characterize the existing tools with respect to all these functionalities, but these concepts could be used (e.g. in the form of a fixed set of keywords) when talking about lexicographic tools, on the EMELD-pages or elsewhere.

Functionalities we identified which are generally not yet covered and therefore are most urgently needed in existing or new tools include:

- Consistency control / management (for different types of consistencies, such as with respect to values of data categories and their ordering) (activities 2 and 3, also affects 4-6).

- the use of morphological parsers specific to linguistic theories or to language types, ideally by plugging in external modules to existing tools (activities 8 and 5).

- Version Control, cooperation (e.g., via the net), logging, registering changes, checking data in and out (activities 4 and 7).

3   Tools Listing

3.1   Comments on format

- The usefulness of ratings is questionable: they are useful only after a sufficiently large number of persons has done an evaluation, and for the software in the EMELD listing, nowhere near that number of people have done ratings (nor does it appear likely they will).
- Another problem with ratings is that they are relative to specific tasks. A tool which someone rated highly with respect to one task might be rated much lower for some other task, even by the same rater.
- Comments (and ratings, if they are kept) should be relative to software version numbers, since bugs are often fixed (and sometimes introduced!), and features added, in newer versions.
- Tools could be indexed by keywords on the EMELD web pages, since some tools cover several functions. It would be useful to know which functionalities of a given program people actually used.
- A more general way to re-organize the tools would be to turn the tools listing website into a wiki.
- Links to other pages that evaluate or discuss tools would be helpful.
- The web pages—however they are eventually structured—will need to be continually updated to reflect new versions of programs (and perhaps programs that are no longer supported).

3.2   Comments on Dictionary / Lexicon Tools Listing

- Some other features that should be discussed or added (cf. Sec 2):
    - Lexical elicitation (e.g. Ron Moe's Dictionary Development Process http://www.sil.org/computing/ddp/), or regional standard word lists, e.g. for comparative linguistics)
    - Interlinear Text (including extraction of new entries from such text)

- Concordance building

- Enriching lexical entries (finding and adding senses, grammatical properties, example sentences, cross-references etc.)

- Consistency control (detecting, fixing, or preventing consistency problems; revision tracking and reverting to earlier versions)

- Export to other formats (XML-based standards, plain text, formats required by morphological parsing tools...)

- Import from standards (XML, text,...)

- Production of presentation formats (e.g. organized by Stem, Root, Citation-Form...; print, web-based or stand-alone electronic formats)

- Embedding media of various types: pictures, sound, video

- Collaboration (network, sneaker-net...; merger with other databases)

- Queries, data mining, retrieval methods

- It would be helpful to have examples of on-line lexical databases, as examples of good practice, evaluated according to criteria such as:
  - Underlying database model

  - Type of information and media provided

  - Off-line version available

  - Dynamic Content Generation (Active Server Pages etc.)

- Standards

  Some participants were skeptical that there could be agreed-on standards for dictionaries/ lexicons, while others felt this was a worthwhile endeavor. Some existing standards (including proposed and *de facto* standards):

  - LMF (= "Lexical Markup Framework"; an ISO proposal, see http://lirics.loria.fr/doc_pub/LMF%20rev9%2015March2006.pdf)

  - OLIF (= "Open Lexicon Interchange Format", see http://www.olif.net/)

  - MDF (= "Multi-Dictionary-Formatter", see http://www.sil.org/computing/shoebox/MDF.html)

3.3   Comments on Individual Lexicon Programs

Not shown here are programs that are mentioned in the EMELD tools listing, but about which the working group did not have any additional comments. All these descriptions could be expanded ad infinitum.

- Shoebox (http://www.sil.org/computing/shoebox/)

  Shoebox was a really great tool in the 90's, but it is behind the times now.

  Pros:

  - Uses plain text databases

  - Interoperability with text/glosses

- Relatively easy to use, like a text editor (but see below re learning curve)

- Has a (crude) morphological analyzer

- Will recognize surface forms of morphs

- Can enter portmanteau morphs

- Flexible: user can create any kind of non-hierarchical field and enter any data in that field. This allows for e.g. page numbers that some reference came from, whether the data needs to be re-checked, etc.

Cons:

- Is trying to parse and disambiguate at the same time, which complicates the issue

- Learning curve is quite steep (training course recommended)

- Uses Item and Arrangement/Item and Process morphology, but does not allow the word and paradigm view

- Insufficient tools for maintaining consistency

- No Unicode (but see below re Toolbox)

- Toolbox (http://www.sil.org/computing/toolbox/)

Toolbox is more or less like Shoebox. perhaps the most important difference is that, it allows the use of Unicode. (Unicode data entry requires—or at least is a lot less painful with—a keyboard remapping program, like Keyman.) It also has a few extra features, such as a 'verify' mode for the morphological parser. Since it is similar to Shoebox, most of the pros and cons for that program pertain to Toolbox as well. Unlike Shoebox, it is still in active development. Freeware.

There is a problem with printing using MDF (but a fix is planned).

- TshwaneLex (http://tshwanedje.com/)

Unicode capable; immediate article preview, customizable fields, automatic cross-reference tracking, automated lemma reversal, online and electronic dictionary modules, export to MS Word format, and network support.

Does not have built-in handling of interlinear text.

- FieldWorks Language Explorer ("FLEX")
  (http://www.sil.org/computing/fieldworks/flex/index.htm)

FieldWorks Language Explorer is a successor to LinguaLinks, also from SIL. It is currently being developed; version 1.0 may be out at the end of 2006, under an open source license.

Language Explorer is organized in 5 areas with a number of "views" of each area (e.g. data entry, publication, etc.). Many views have both a browse pane (minimizable) and an edit pane (minimizable). The areas (not all of which will be feature complete in the first release) are:

- Lexicon (many features, e.g. concordance view for defining senses)

- Interlinear Text (includes a built-in morphological parser)

- Grammar (currently morphology; syntax later)

- Integration (with other modules of FieldWorks, such as a translation editor)

- Script Handling (Unicode, including complex non-standard scripts)

Pros:

- Makes data consistency easy

- Intended to guide user in e.g. morphological discovery (how well this will work remains to be seen)

Cons:

- Inflexible, e.g. adding user-defined fields is highly restricted.

- Highly computer intensive, i.e. lots of memory and fast CPU

- like Shoebox/Toolbox, seems to impose a certain model of linguistic reasoning on the user; applicability for word-and-paradigm approaches has still to be demonstrated

● LEXUS (http://www.mpi.nl/lexus/)

LEXUS is being developed at the Max-Planck-Institute in Nijmegen. It is based on the LMF data model (the proposed ISO-standard), and thus very flexible with respect to different data categories and their ordering, but yet will allow to compare and combine different lexical databases. Being sort of a lexical complement to ELAN (for annotation of audio and video data), LEXUS allows for inclusion of all sort of multi-media material and for defining relations of all kinds between entries or data categories or their parts, and to other external resources such as (parts of) texts.

Lexus is implementing templates to help the naive user work with the dictionary tool. it will be open source. Currently only an on-line version is available (in particular, the user interface is still in an alpha stage).

● Shoeshonedictionary (http://shoshonedictionary.com)

This dictionary was mentioned in the working group, but it is unclear to us if they provide technology that would allow to be applied to other languages, such as an SQL dictionary.

● Filemaker Pro (http://www.filemaker.com/)

A relational database that several sophisticated users have used to build lexical databases for particular projects.

It has problems past 4,000-something records; compresses things together and is a hardware-related tool.

Is a general tool, not specially designed for lexicography, and therefore does not allow certain procedures, at least not in a straight-forward way. Ex. Cannot gloss by just using overt morphs: "Go." imperative singular/plural, etc. (These comments were mentioned by participants, but we are not sure to what exactly they refer.

● Lexique Pro (http://www.lexiquepro.com/)

A program for displaying and distributing lexicographical databases originally in standard format (Toolbox). Basic editing is possible, but this it not a program meant for building dictionaries. Converts databases (including pictures and other multi-media resources) into a standalone electronic or WEB version with automatic hyper-links etc.