Multilingual Lexicon-Text Database

Database structure

Draft, 03.09.2001

Christian Lehmann
University of Erfurt

1. Introduction

1.1. Functionality

The aim of the project ‘Multimediale Lernmaterialien für Japanisch und Chinesisch’ is to develop learning materials for speakers of German and English who learn Mandarin Chinese or Japanese. The materials consist in a corpus and a lexicon. The corpus contains a collection of texts each of which embodies a scene or story which is matched by a video and audio recording. Thus, a text is represented at various linguistic and medial levels. The lexicon consists of a set of entries each of which contains linguistic and cultural information at various linguistic levels and in various media.

The entries of the lexicon may have various degrees of complexity. Thus, not only words, but also idioms and proverbs are possible lexical entries. Therefore, the lexicon will simply be called ‘inventory’ in what follows. A complex entry of the inventory has an internal morphological and/or syntactic structure. The entries of the inventory are related among each other by various conceptual and structural relations.

Each token appearing in the text corpus is either an instance of an entry of the lexicon or composed of tokens for which this is true. Complex text tokens have an internal morphological and/or syntactic structure. Complex inventory entries and complex text tokens alike each instantiate a schema. Schemata are items of the inventory and thus have the same status as lexical entries. Insofar, lexicon and grammar are treated alike, as two parts of the inventory which mainly differ in their degree of generality or abstractness.

The software developed in the project has the following functionality at the front-end:

- The user language and the language to be studied may be chosen.
- A lesson may be retrieved by various selected criteria.
- A lesson may be played as a video including an audio track which in turn is accompanied by captions.
- The text of a lesson may be represented at various levels which may be combined freely.
- The grammatical structure of a corpus sentence may be visualized.
- For a given text token, the inventory entry instantiated by it may be retrieved.
- For a given inventory entry, selected kinds of information may be displayed.
- Navigation in the inventory along the relations among entries is possible.

S Inventory entries may be grouped by various selected properties.
S For a given inventory entry, the sentences which contain tokens of it may be displayed.

1.2. General implementation

The multimedial learning software is accessible either via the internet or from a DVD. To provide for the user-functionality described in the previous section, the software consists of three main sections, a front-end, a back-end and an interface between them.

At the back-end, the data of the inventory and the corpus are stored in a relational database. Technically speaking, what the user sees and hears is a report from the database.

At the front-end, the user may navigate in the database report and manipulate it to trigger new reports. The front-end is implemented as a plug-in of a WWW browser. It may retrieve the data either via the internet or from a DVD.

The interface between the front-end and the back-end sends the user’s requests to the database and sends the database reports to the user. It may be implemented in SQL.

2. The structure of the database

The database at the back-end bears the name of ‘Multilingual Lexicon-Text Database’ (MLTD). In the present conception, the four languages Chinese, English, German and Japanese are logically equal, although language-specific information may be elaborated to varying degrees. Thus, the MLTD may, in principle, also form the basis of a German-learning software for speakers of Japanese.

The MLTD consists of a number of related tables which are described below. For each database table, the record structure is specified by a table in the following text which enumerates the fields. The text tables have the following structure:

Column 1: field number
column 2: field name,
column 3: characterization of field content,
column 4: data type of field.

Table 1.	**Concepts**	Master list of concepts	
1	Concept ID		
2	Concept Name	English word or phrase	text
3	Definition	(informal) definition; optional	text
4	Visual representation	for concrete, culture-specific concepts; optional	JPEG or MPEG file

The concepts may be language-independent or language-specific. They are used to associate expressions of one language with semantically related expressions of the same language or with equivalent expressions of another language. Thus, what the user sees is not concepts, but expressions (via table 5).

Fields 2 and 3 are chiefly used by system administrators. If the user wants to see the English expression for a given concept, table 5 (for English) is used.

Table 2.	**Conceptual relations**	Master list of conceptual relations	
1	Relation ID		
2	Relation Name	is a (kind of), is part of, is property/aspect of, is cross-related with	text

The conceptual relations are defined in terms of their logical properties in Lehmann 1996.

Table 3.	**Conceptual network**	Relations among concepts	
1	Concept 1	first related concept: Concept ID	link to table 1
2	Relation	Relation ID	link to table 2
3	Concept 2	second related concept: Concept ID	link to table 1

The conceptual network is language-independent. The relations are many-to-many. Among other things, the network provides a taxonomy (relating hyperonomyns and hyponyms to a given concept) and a meronomy (relating parts to wholes).

Table 4.	**Inventory**		
1	Entry ID		
2	Entry Name	for grammatical categories: English abbreviation; for individual words: void	text
3	Category	immediate supercategory: Entry ID	link to table 4
4	Number of operators	zero, one	byte
5	Number of operands	zero, one, two	byte

6	Operator	Entry ID	link to table 4
7	Operand 1	Entry ID	link to table 4
8	Operand 2	Entry ID	link to table 4
9	Adjacency	construction is (or is not) obligatorily continuous	boolean
10	Operator Position	preceding first operand, following first operand, either side of first operand, following second operand	byte
11	Orthographic Representation I	Sequence of Unicode Characters (Kanji)	text
12	Orthographic Representation II	Sequence of Unicode Characters (Kana)	text
13	Alphabetic Representation	Sequence of Unicode Characters (ASCII)	text
14	Phonological Representation	Sequence of Unicode Characters (IPA)	text
15	Phonetic Representation	native pronunciation	MP3 file
16	Morphological classes	[to be elaborated]	
17	Register	[to be elaborated]	text
18	Historical stage	[to be elaborated]	text
19	Other	[to be elaborated]	

The entries of this table are language-specific, i.e. they belong to one of the languages of the MLTD. There is one table like table 4 for each of the languages[1]. All links from other tables to inventory entries (i.e. to field 1) have to refer to the language-specific table.

The entries of table 4 may be simple or complex. The simplest entries are morphemes (roots and grammemes); the most complex entries may be sentences such as proverbs.

An entry of the inventory may be more or less schematic. That is, it may be a pure construction, or the whole entry or parts of it may be individual signs.

[1] Alternatively, there could be only table 4 for all of the language. In that case, table 4 would need an additional field specifying the language.

A construction is a schema in two senses: It is a schema for other entries of the inventory, and it is a schema for complex tokens of table 7.

Schematic (taxonomic) relations between entries are represented in field 3. For the uppermost category of a taxonomic hierarchy (the “maximal projection” of a syntactic category), it may be appropriate to repeat the entry ID in this field.

Schemata are underspecified. If x is a schema for y, then there is an inheritance relation such that x inherits the specifications of y.

Polysemy, allomorphy and morphological irregularity are schematic relations. I.e., an irregular form is an entry which is an instance of a more general entry. The former refers to the latter in field 3.

Meronomic relations between entries are represented in fields 4 – 8. If the fields are empty, the entry may or may not be complex. For non-branching categories, it may be appropriate to specify the number of operators as 0 and the number of operands as 1.

Conversion of category is represented as a meronomic relation to one operand, with zero operator.

Field 13 contains a Romazi representation for Japanese and a tonal representation for Chinese.

Fields 2 and 13 are complementary in that if one of them contains information, the other does not. However, they are not fused because the emptyness of a field may be used to technically identify the kind of record.

Table 5.	**Meanings of inventory entries**	Relations between concepts and language-specific units	
1	Concept ID		link to table 1
2	Inventory entry	Entry ID	link to table 4
3	Conditions	context conditioning the reading ‘Concept ID’: position in construction	link to table 4

Table 4 does not contain semantic information. Since one expression (inventory entry) may signify different concepts (polysemy) and one concept may be expressed by different inventory entries (synonymy), relations between concepts and inventory entries are many-to-many and require a separate table.

The exact form of the content of field 3 remains to be elaborated. Probably a complex reference which combines a link to field 1 of table 4 with one of the numbers (of fields) 6 – 8 is appropriate to identify the position of the item in a construction.

Table 6.	**Scenes**	Inventory of the scenes of the corpus	
1	Scene ID		
2	Lesson Number	number of the lesson in the text corpus	integer
3	Scene Number	sequential position of scene in lesson	integer
4	Video clip	scene with sentence spoken	MPEG file

Table 6 only contains scenes in an ordered sequence. It does not contain the sentences themselves, but only their sequential order in the corpus. A sentence is composed of tokens of table 7, which refers to field 1 of table 6.

From table 6, a lesson may be output as a sequence of video clips.

In the present conception, it is assumed that scenes are culture-specific, and therefore they include an audio file. Alternatively, audio files could be associated with text tokens (sentences) of table 7.

Table 7.	**Text tokens**	Simple and complex units of a text	
1	Token ID		
2	Scene ID		link to table 6
3	Sequential Number	sequential position of token in sentence	integer
4	Inventory unit	Entry ID of inventory	link to table 4
5	Operator	Token ID	link to table 7
6	Operand 1	Token ID	link to table 7
7	Operand 2	Token ID	link to table 7

The scenes of table 6 are coupled with text tokens. A text token is an instance of an inventory entry (cf. Lehmann 1998 for the linguistic basis of this conception). I.e., technically the text corpus of MLTD does not consist of text, but of links to table 4. A token consists of zero up to three other tokens. The number of constituents is determined by fields 4 and 5 of the corresponding entry of table 4. Accordingly, fields 5 – 7 may be empty.

Terminal (non-compound) tokens are those whose inventory unit is void in its field 2. For these, only fields 1 – 4 of the present table are used.

For non-terminal tokens, fields 3 and, possibly, 2 are not used.

For each language of the MLTD, there is one table like table 7. Tokens from different tables may refer to the same scene of table 6. In this way, a translation for a sentence may be provided.

3. Illustration

The following table is a segment of table 4 for English. Record numbers are arbitrary.

Example entries of English inventory (table 4)

1	Entry ID	101	111	102	103	112	113	104	105	106	107	108	109	110
2	Entry name					Adj	AdjP	N_{comm}	DetNom	Art	Det	Nom	NP	Pers. Pron.
3	Category	106	112	104	105	113	113	108	109	107	107	108	109	109
4	Number of operators				1				1	0	0	0		
5	Number of operands				1				1	1	1	1		
6	Operator				111				107					
7	Operand 1				102				108	106				
8	Operand 2													
9	Adjacency								+					
10	Operator Position								1					
11	Orthographic Representation I													
12	Orthographic Representation II													
13	Alphabetic Representation	the	supreme	court	Supreme Court									
14	Phonological Representation	ði	səˈpɹim	kɔɽt										
15	Phonetic Representation													
16	Morphological classes													

17	Register	-	formal	-	-									
18	Historical stage	modern		modern	modern									
19	other kinds of information													

References

Lehmann, Christian 1996, "Linguistische Terminologie als relationales Netz”. Knobloch, Clemens & Schaeder, Burkhard (eds.), *Nomination - fachsprachlich und gemeinsprachlich.* Opladen: Westdeutscher Verlag; 215-267.

Lehmann, Christian 1998, "Programme de description globale d'une langue (Language Description System)." *Lingua Posnaniensis* 40:103-124.