CLIPP

Christiani Lehmanni inedita, publicanda, publicata

titulus
Structure of a comprehensive presentation of a language
huius textus situs retis mundialis
http://www.uni-erfurt.de/sprachwissenschaft/personal/lehmann/CL_Publ/interfac.pdf
dies manuscripti postremum modificati
28.01.2002
occasio orationis habitae
University of Tokyo, 21.11.2001
volumen publicationem continens
Tsunoda, Tasaku (ed.), *Basic materials in minority languages 2002*. Osaka: Osaka Gakuin University (ELPR Publications Series B003)
annus publicationis
2002
paginae
5-33

Structure of a comprehensive presentation of a language

With particular reference to the interface between text, grammar and lexicon

Christian Lehmann

University of Erfurt

Christian.Lehmann@Uni-Erfurt.De

Abstract

The comprehensive presentation of a language properly includes its documentation and its description. The documentation contains a text corpus whose texts are represented at various linguistic levels. The description properly includes the grammar and lexicon of the language.

All of these components are interrelated in a systematic fashion. If the presentation of the language is implemented in the form of a relational database, the in-built relationality of the database is the ideal way of representing those relations that are systematic. Other cross-references may be implemented in the form of hyperlinks. This is demonstrated in most detail for the lexicon.

1 Introduction

The purpose of this contribution is a methodological one: The structure of the comprehensive presentation of a language is articulated in such a way that it may be implemented in the form of a database.[1] This requires a subdivision of the overall presentation into components which are separate, but interrelated. Particular attention will be paid to the interrelation, or interface, between the text corpus, the grammar and the lexicon. The structure is neutral as against particular languages, so the database may house any language to be documented and

[1] The research reported here was partly supported by grant Le 358/8 of Deutsche Forschungsgemeinschaft.– On methodological aspects of language documentation, more in general, see Tsunoda 2001.

described. What is proposed here is in many respects a maximum model. Several specific kinds of data or of representations may not be needed in the presentation of a particular language.

Given its object and goal, the paper pertains to the following linguistic disciplines:

- grammaticography, the methodology of grammar writing,
- lexicography, the methodology of lexicon elaboration,
- text edition, the methodology of preparing a text corpus.

It is assumed that all of these activities are computer-aided. The paper presupposes a theory of linguistic description and entails a set of instructions to the analyst telling him how to go about his work. Insofar, the paper has a practical purpose.[1]

1.1 Documentation and description of a language

The distinction between the documentation and the description of a language pertains to the methodological level:

A **documentation** of a language is a collection of primary linguistic data in the form of a text corpus (including ideally a correlated corpus of video or audio recordings), represented at various linguistic levels and possibly with annotations. Its object is speech (*parole*) rather than language (*langue*). Its structure is determined by the sequence and structure of the texts that it consists of.

A **description** of a language is an account of the system underlying the (documented) data. Its object is language rather than speech. Its structure is determined by the nature of human language and its components, by the individual language system and by the methodological perspectives taken on it.

A description is, thus, at a meta-level with respect to a documentation (cf. Lehmann 2001, section 4). Since documentation and description of a language should always go together, a term is needed for the union set of both. For such a comprehensive account of a language, the term **presentation** of a language will be used.

---

[1] The basic ideas of the system proposed here were already set out in Lehmann 1998, which is freely made use of in what follows.

1.2 Relational database

As every linguist is aware, the language system consists of units which belong to classes which are connected by relations. These form a multidimensional space. The presentation of a language in the form of a book can hope to bring out the hierarchy – either the taxonomy or the meronomy – among its components, but it cannot possibly represent adequately the multidimensional relational structure. This can be achieved if the presentation is implemented as a relational database. Such a representation is not only adequate to the nature of the object, but is also of great methodological value, since it helps and forces the analyst to keep his categories and relations consistent.

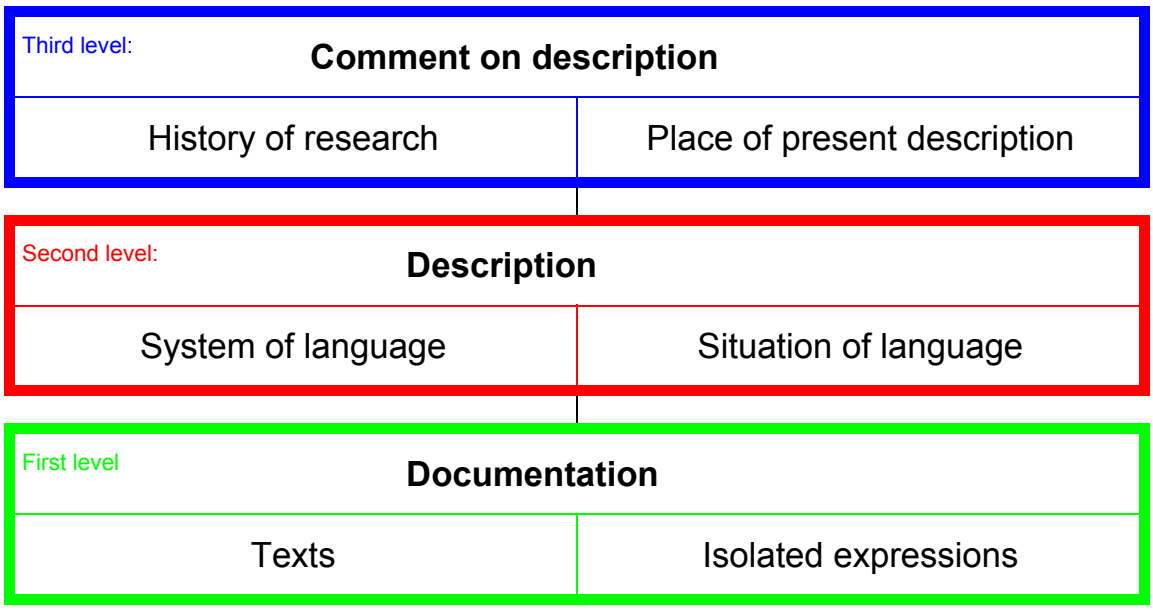
The technical structure of the relational database will not be specified in any great detail in the following account. For instance, when two components of the description cross-refer to each other, the first one will mention the second one, and vice versa. The technical solution is a bit simpler, but for the sake of better understanding, the redundant way of cross-referencing will be explained here.

2 Global structure of the presentation of a language

2.1 Components of the presentation

The comprehensive presentation of a language takes the global form of S1.

S1. *Presentation of a language*



In S1, three methodological levels are distinguished: At the level of the object itself, we have the **documentation** of the language proper in the sense defined in section 1.1, viz. a corpus of texts, video and audio files. At the second level, we have the **description** of the language. This level is related as a meta-level to the first level. At the third level, we have the **comment** on the description, where the analyst accounts for the conditions of and decisions taken in his documentation and description. This is again a meta-level with respect to the second (and first) level.

Apart from these three core components, a full presentation of a language contains further material, including, in particular, a bibliography in the form of a formatted set of references. On the other hand, some additional sections that we are familiar with in scientific publications, such as indices of terms, are not needed, as they are inherent in the structure of the model, viz. the relational database.

At the level of the description of the language, the main distinction to be made is between the situation (or setting) and the system of the language. While the former will be dealt with briefly in section 3, the bulk of this paper will concentrate on the **language system**, whose structure is displayed in S2.

S2. *System of language*

<table>
<tr><td colspan="4">Level 2: Description
System of language</td></tr>
<tr><td colspan="2">Semantic system</td><td colspan="2">Expression systems</td></tr>
<tr><td>Lexicon</td><td>Grammar</td><td>Primary:
Phonology</td><td>Secondary:
Writing</td></tr>
</table>

We can see that in the terminology adopted here, phonology is not part of grammar, but belongs to the expression systems of the language. Naturally, the expression systems, too, are systematically referred to in the other components of the presentation. For instance, morphological rules contained in the grammar refer to phoneme classes defined in the phonology. Although such references do play a role in what follows, the expression systems will be presupposed in the present account.

From this it becomes clear that the three components of the presentation of a language which are mentioned in the subtitle of this paper do not constitute a

proper part of it. Instead, texts belong to one level, the object level, while grammar and lexicon belong to another level, the level of the description of the object. Furthermore, lexicon and grammar, in turn, do not exhaust the components situated at this level, and instead form a proper subsystem of the description.

From the point of view of structural linguistics, these three components nevertheless bear systematic relations to each other (cf. Seiler 1969). The syntactic and morphological categories and word classes of the grammar form a **taxonomy** in which a unit of a lower level is an instance of a unit of a higher level, which serves as a schema for the former. For instance, a proper name, N_{prop}, is an NP, and it instantiates the schema of the NP. This taxonomy extends into the lexicon, whose items form its bottom. Thus, e.g., *Henry* instantiates the schema of N_{prop} (cf. Langacker 1987, ch. 6).

At the same time, morphs as they occur in texts are tokens of morphs which are allomorphs of morphemes of the lexicon. This is, thus, a **type-token** relationship. These two kinds of relations between units at the various levels are summarized in T1.

T1. *Logical relations between units of grammar, lexicon and text*

component	unit	relation	
grammar	word class	schema	
lexicon	word = lexical entry	instance	type
text	word = text occurrence		token

2.2 Implementation in a database

The model sketched above is implemented in the form of a relational database. Other options are conceivable. In particular, many of the relations between items contained in the presentation can be implemented by links in a **hypertext**. These are appropriate for ad-hoc references to other items or passages in the presentation. When it comes to systematic relations, for instance the taxonomy of grammatical categories or the type-token relationship between lexical items and occurrences in the text corpus schematized in T1, hypertext links are both impractical and insufficient. They are impractical because a decision which has to be made only once, at the level of the design of the interface of the main components, would have to be made separately for each item involved. And they

are insufficient because **logical properties of relations** among elements cannot be controlled or used for consistency checks. For instance, the hyponymy relation and the part-whole relationship are transitive.[2] This allows the analyst to erect a taxonomy and a meronomy over the concepts involved and check the consistency of these hierarchies by an algorithm. This is possible if hyponymy and part-whole relation are among the relations implemented in the database; it is impossible if they are implemented as hyperlinks from each given concept to its particular hyperonym or whole, respectively.

Theoretically as well as technically, the schema-instance relationship (as well as the part-whole relationship) differs from the type-token relationship in T1. In terms of the latter contrast, both schemas and instances are types. The technical implementation of this relationship is therefore in terms of a link between two records of the database. Contrariwise, in the latter relationship, the tokens are, technically speaking, instant copies of a type. In concrete terms, this would entail that the texts of the documentation would be generated from the language system in the moment in which they are used. Since this is not feasible, the type-token relationship must be added to the set of logically different relations distinguished in the structure of the database. Thus, this relationship, too, is implemented as a link between two records, but links are labeled for the kind of relation they represent.

A relational database requires a high amount of standardization and of uniformization. For instance, the lexicon will contain native and loan words. For loan words, there must be a field in the table which contains the donor language of the loan, a field which is superfluous for the bulk of native items. Once a field has been set up, a relational database requires it to be there even in records where it makes no sense. However, on the one hand, this rigidity helps the analyst in clarifying his thought, in controlling the consistency of his account and in automatically treating in an analogous fashion all those many cases which are, in fact, analogous in the language under study. On the other hand, a relational database does have the necessary flexibility to adapt itself to inconsistency and variation in the object to be described. In particular, in addition to the relational potential built into the database structure, hyperlinks may freely be made use of.

The diversity in the kind of data and representations contained in the presentation is also reflected in the diversity of **data types** assigned to the fields of the database. Some of the fields have free text ("memo") format. Others have more

---

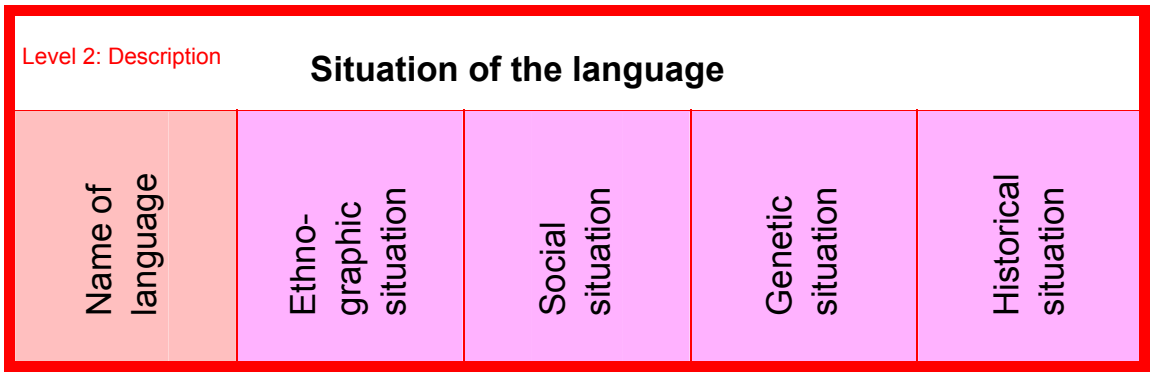
[2] If x is a hyponym of y, and y is a hyponym of z, then x is a hyponym of z.

formal data types. Many contain an item from a **range set**. This means that the permitted contents of a given field are taken from a closed set. For instance, the field 'part of speech' in the entries of the lexicon can only contain one of a limited number of concepts.

3 Setting of the language

While the system of the language is, so to speak, the language from within, the setting of the language refers to the language viewed as a whole in its context. The subdivision of the account of the situation of the language is displayed in S3 (cf. Bohnemeyer et al. 1994 for more detail).

S3. *Situation of the language*



First, the **name of the language** needs clarification, as the language will be known under different names, all of which have their history and use. The information in the section on the **ethnographic situation** of the language is, in principle, the subject matter of ethnology, not of linguistics, and would, consequently, be fully spelt out in an ethnographic description of the community or society speaking the language. It will be mentioned in a description of the language only to the extent that it is useful for the understanding of texts and examples. This includes information on the geographical distribution of the language, the ethnic affiliation of the speech community, the organization of the society and its culture.

The **social situation** of a language has an internal aspect, which relates to its speakers, its stratification and the existent communicative conventions. Its external aspect relates to the competing languages and the status of the object language.

The **genetic situation** of the language, again, has an internal aspect, which is its subdivision in terms of dialects, and an external aspect, which is its genetic

affiliation. The same subdivision applies, for the third time, to the **historical situation** of the language. The external history concerns migrations and contacts of the speech community and the development of writing and literature, while the internal history is the development of the language system from the proto-language to the modern state.

The other components of the presentation refer extensively to this component. In particular, the genres of texts contained in the text corpus (cf. T3 below, field #13) are accounted for systematically in the ethnographic situation of the language.

4 Lexicon

4.1 The lexicon as a relational database

The lexicon of a language consists of several thousands of lexical items.[3] It is easy to conceive of them as the entries of a database. The various linguistic properties of a lexical item, such as its phonological form, its meaning, its grammatical category and so on, will then constitute the fields which compose each entry (and the columns of a database table). We will see in a moment how many of such properties or fields there are, in fact. What tends to be forgotten is the relational nature of the information in the lexicon. A lexicon is not just a set or an inventory of items. Items fall into multiple classes, some of which form hierarchies inside the lexicon, while others transcend the boundaries of the lexicon. And items bear multiple lexical relations to each other. For instance, the fact that Engl. *hot* is the antonym of *cold* is not an absolute property of the lexical entry *hot*, but instead a relation which connects the two entries *hot* and *cold* just as it connects dozens of other entries. In this way, the lexicon is a relational network rather than an inventory of entries.

The most wide-spread software tool for the confection of lexical databases is Shoebox™.[4] Much of what a linguist requires of a lexical database program is provided by Shoebox,[5] and it has many features such as user-defined sort orders which all other database programs on the market lack. Those readers familiar

---

[3] The number of morphemes of a language ranges approximately between a few thousands and ten thousand. This is also the size of the average lexicon of an "exotic" language. The number of lexemes in a lexicon of a language of "civilization" can be much higher.

[4] In 2001, version 5 is available from SIL.

[5] Cf. Wimbish & Davis 1992, the handbook accompanying an earlier version of Shoebox, which is, at the same time, a highly useful introduction to computer-aided lexicography.

with Shoebox will recognize many of the concepts used in what follows. However, Shoebox has no relational potential[6] and insofar does not mirror, at the level of the scientific representation, the intrinsic structure of the object.

4.2 Structure of a lexical entry

In what follows, the structure of the lexicon will be presented as the structure of a lexical entry. In a database, primary order of entries is no issue of relevance. Instead, entries may be sorted according to the content of any field desired, not only the lemma itself, but also the semantic class, the English translation or any other property. T2 is a comprehensive list of properties of a lexical entry that may be relevant in the lexical component of the presentation of a language.[7] The one-but-last column of the table contains examples for selected lemmas, in the form 'content of field #1: content of field of this row'.[8] The last column of T2 contains the components of the overall system which the database field in question refers to.

---

[6] The program partially compensates for this by its 'jump path' feature.

[7] Some of the examples given below are taken from the database of the Yucatec Maya lexicon that we are elaborating at the University of Erfurt.

[8] Since probably no lemma of no language requires the full set of fields, the examples are taken from various lemmas of various languages.

T2. *Structure of a lexical entry*

category	**nº**	**field name**	**explanation**	**example**	**related comp.**
entry identity	1	lemma	standard orthographic representation	Lat. *ago*	
	2	homonym number	number in a series of homonyms	Engl. *down*: 2	
	3	citation form	form in which the lemma is mentioned	Lat. *ago*: *agere*	
	4	mother entry	polysemous entry of which current entry is a sense [link]	Engl. *paragraph:* ú*paragraph*$_0$	lex.
expression	5.a	phonetic representation	segmental/prosodic representation (IPA)	Engl. *egret*: [i+g bt]	phon.
	5.b	sound	acoustic recording [sound file]		
	6	orthographic variants	orthographic representation	Engl. *neighbor*: Brit. *neighbour*	writing
	7	phonological variants	phonemic representation	Engl. *egret*: /igret/	phon.
language	8.a	dialect	regional variety	Engl. *buck*$_2$: American	gen.sit.
	8.b	sociolect	sociolect or special language	Engl. *nominalize*: scientific	soc. sit.

	8.c	style	style or register	Engl. *buck*$_2$: colloquial	soc. sit.
	8.d	stage	diachronic language stage	Engl. *bereave*: obsolete	hist. sit.
structure	9	proper name	linguistic name of a grammeme	Engl. *-ly*: adverbializer	gr.
	10	syntactic category	word class, subcategory	Engl. *butter*: mass noun	gr.
	11.a	morphological structure	immediate morphological constituents [links]	Engl. *nominalize*: ú*nominal*, ú*-ize*	lex.
	11.b	word formation	last derivational process applied	Engl. *nominalize*: verbalization	gr.
	11.c	derivatives	set of lemmas of derived records [links]	Engl. *nominal*: ú*nominalize*	lex.
	12	morphological categories	set of morphological categories	gender/noun class, possessive class, inflection class	gr.
	13	irregular inflection	(paradigm of) irregular inflected forms	Engl. *good*: comp. *better*	gr.
	14	construction	distribution, selection restrictions	Engl. *depend*: [_ [on [X]$_{NP}$]$_{PP}$]$_{VP}$	gr.
	15	phraseology	collocations, phrases etc.	Engl. *approach* (n.): *take an _ to X* = *úapproach* (v.) X	gr.

meaning	16	native definition	semantic explanation in object language	Yuc. *utskint*: *kin wutskintik: kin bèetik u yutstal*	
	17.a	meaning 1	semantic explanation in English	Yuc. *utskint*: make good, enhance, repair, cure	
	17.b	meaning 2	semantic explanation in regional language [native translation]	Yuc. *utskint*: mejorar, componer, curar	
	17.c	meaning 3	semantic explanation in background language	Yuc. *utskint*: gut machen, bessern, reparieren, heilen	
	18.a	gloss 1	interlinear morpheme gloss in English	Yuc. *utskint*: good:FACT	
	18.b	gloss 2	interlinear morpheme gloss in regional language	Yuc. *utskint*: bueno:FACT	
	18.c	gloss 3	interlinear morpheme gloss in back-ground language	Yuc. *utskint*: gut:FAKT	
	19	semantic classes	set of classificatory features	Engl. *deer*: animal	lex.
	20	semantic relations	set of lexical relations to other records	Engl. *stag*: hyperonym: ú*deer*, immediate contrast: ú*hind*	lex.

	21.a	encyclopedic information	ethnographic description	Engl. *stag*: Linné *cervida*. Bigger than buck. Is hunted and eaten. Antler is a demanded trophy.	ethn. sit.
	21.b	picture	visual image (image/video file)	Engl. *trefoil*: Ê	ethn.sit.
	22.a	origin	native, or donor language of loan	Engl. *trefoil*: French	soc. sit.
	22.b	etymology	for native: original expression/structure and meaning	Engl. *hussy*: *housewife*	hist. sit.
	22.c	cognates	for native: set of cognate expressions from other languages	Engl. *deer*: Germ. *Tier* 'animal'	gen.sit.
methodology	23	comment	additional, esp. methodological information	Jap. *sugi*: Informants from Kyoto and Tokyo agree on the pronunciation [s , õi], not [s , gi].	meth.
	24	problems	as yet unsolved problems	Yuc. *t'úul*: verify high tone	meth.
	25	date	date of last modification (automatic)	17.11.2001	

Some of the fields of this lexical entry are self-explanatory. In what follows, I will comment on the other fields.

Homonym number: Homonyms are, of course, separate entries distinguished by numbers. The same goes for the readings of a polysemous entry. See below section 4.3.

In the 'structure' subsection of the lexical entry, which now follows, the concepts used are those introduced systematically in the grammar, to which reference is implied here.

Proper Name: What is meant here is the proper name of a grammatical morpheme. For instance, the proper name of the English suffix *-ize* is *verbalizer*. Consequently, the possible contents of this field are unique, and only a minor portion of the entries of the lexical database will be specified for this field.

Morphological structure: This field indicates the internal morphological structure of the lemma by referring to its immediate constituents, which are entries of the same lexicon. The latter refer back to the present entry by virtue of their field #11.c, for which see below.

Word formation: Possible entries in this field are taken from a range set defined in the grammar, where the word-formation processes of the language are dealt with systematically.[9] In this field, the last word formation process applied is indicated, i.e. the process which was applied to the components of field #11.a to form the stem of the lemma. In the case of a derivationally complex lemma, other word formation processes may have created stems that are part of it, in particular those of field #11.a. Such processes are not indicated here, since they may be seen by following the links of the latter field.

Derivatives: This field refers to existing derivations. Since they must constitute independent entries of the lexicon, these references are just links to other entries of the database. This field is in a mirror relation with field #11.a.

Syntactic category: From among the grammatical categories of a lexical item, this field is dedicated to its syntactic category qua distributional category (for morphological categories see #12). This is understood as a narrow subcategory of a part of speech, e.g. 'transitive verb with additional prepositional

---

[9] To give an example of what the content of this field looks like: in the Yucatec Maya lexical database, the word formation processes include adjectivization, (basic,) causative, compound, deagentive., denominal, deverbal, distributive, durative, extroversive, factitive, fin_verb_form, incorporative, intensive, introversive, passive, phrasal, positional, processive, reduplicative, reflexive, spontaneous, totally_affected, usative.

complement'. A lemma which belongs to diverse syntactic categories is considered polysemous. Each category then constitutes a record (see below section 4.3).

Construction: This field shows the syntactic and semantic construction frame of the lemma, including selection restrictions. This is a specification of the information contained in the previous field. It should be represented by a formal notation, e.g. '_ X' (_ = position of lemma, X = relevant syntactic constituents or properties of the context).

Morphological categories: An inflecting word of a language may fall into diverse morphological categories at once. Some may be syntactically relevant lexical classes such as the gender of a noun, others may be purely morphological classes such as inflection classes. It is practical to set up a separate field for each of these categories.

Irregular inflection: Either the stem or the inflectional category or both may have suppletive expression. Such inflected forms which are not derivable by rule are enumerated here, e.g. '3rd pers.sg.: *has*' in the lemma of Engl. *have*.

For the following fields, a couple of metalanguages or background languages come into play. The **meaning** of the lemma is first explained in the object language being described (field #16), as it would be the case in a monolingual dictionary. Next (field #17), the meaning is specified in the regional language, because that may be the language in which the linguist and the informant cooperate and in which the dictionary may be published. The native translation is a translation provided by the author of the lemma or another speaker of the object language. For practical purposes, the meaning is also given in the native language of the analyst. Finally, the meaning is indicated in the language of publication, which more often than not will be English. The meaning is specified in plain prose. What is easily formalizable about it is relegated to other fields of the lexical entry, in particular 19 and 20.

Gloss: Each lemma has a unique gloss which represents it in interlinear morphemic glossing. Just as in the case of the meaning of the lemma, there is a set of glosses from different metalanguages. Each gloss consists of a sequence of morphemes or category labels taken from the respective metalanguage. Their format is standardized; cf. Lehmann 1982.

Semantic classes: Every lexeme falls into one or more semantic classes. For instance, Engl. *apple* belongs to the classes 'fruit' and 'food'.[10] These classes

[10] To give an impression of what kind of semantic classes may be used, here is the list used in the Yucatec Maya lexicon: abstract_property, animal, artefact, behavior, behavioral_property, bird, body_liquid, body_part, celestial_body, cognition, color, communication, contact, disease,

form another range set and constitute lexical fields. Depending on the language, some of them may be grammatically relevant and, thus, reappear in the grammar.

Semantic relations: Here the lexical entry is connected with other lemmas by paradigmatic lexical relations like synonymy, antonymy, hyponymy, hyperonymy, cohyponymy. As mentioned in section 2.2, hyponymy and the part-whole relation lead to a taxonomy and a meronomy, respectively.

Encyclopedic information: The content of this field goes beyond linguistic semantics, giving information on real-world, especially culture-specific properties of things designated. This field may refer to the pertinent section of the 'Situation of the language', esp. the ethnographic situation, for background information.

Usage: This concerns style, register, connotations and any kind of pragmatic information.

Etymology: This field has the double function of indicating the reconstructed base for native words and the origin for loans.

Cognates: This field contains cognate, i.e. formally or semantically related, words from genetically related languages.

Comment: This field contains any additional information, esp. of a methodological, stylistic, sociolinguistic nature, including the status of the lemma and ungrammatical examples.

Problems: This field contains questions to be investigated and problems to be solved in future lexicographic work, especially fieldwork. This field is directly related to the previous one in that the latter contains the solutions to problems that had been formulated in the present field.

Date: This is the date of last modification, which the database program will update automatically.

4.3 Relations among entries

Some fields have a classificatory function. For instance, permitted entries of the field 'semantic class' are standardized. It is thus possible to select and display

---

durative_action, durative_process, emotion, emotional_behavior, emotional_expression, evaluation, evaluative_property, evidential, food, gesture, illness, iterative_action, iterative_process, kin, local_state, motion, perception, perceptual_quality, person, phase, physical_action, physical_process, physical_property, physical_sensation, physical_state, place, plant, plant_part, posture, psychic_quality, psychic_state, psychosomatic_process, psychosomatic_state, punctual_action, punctual_process, social, sound, spatial_region, substance, terminative_action, terminative_process, time, volition, weather_condition.

simultaneously all the lexical members of one semantic class. The same goes for the grammatical category of lexical entries.

Links among lemmas are based on the content of a particular field of their records. For example, a derived base contains, in its field #11, the stems or morphemes which form it and which constitute lexical entries by themselves. Technically, this is a link from a section of the content of field #11 of one record to field #1 of another lexical record. Similarly, the content of such fields as 'semantic relations', 'derivatives' etc. boils down to links with other lexical entries.

Finally, there are two notorious problems of lexical semantics and lexicography: How does one deal with the distinction between **homonymy and polysemy**; and how does one represent the related senses ("readings") of a polysemous item? Both problems find an elegant and satisfactory solution in a lexicon which is implemented as a relational database. The set of senses of a polysemous word is not contained in field #17 of one lexical entry. Instead, there is an independent entry for each such sense. These entries, of course, have their lemma in common, but differ in the content of their field #17. However, this need not be their only difference. Polysemy is often related to other lexical differences, for instance in construction, in stylistic status and so on. This poses a problem of representation if the senses are enumerated in one lexical entry. If they constitute distinct entries, they each receive their homonym number (field #2) and are, insofar, treated like homonyms. The difference between homonymy and polysemy resides in the fact that the entry of a sense of a polysemous word refers to a mother entry (field #4) which represents the polysemous item, while there is no such link among homonyms. This intuitively reflects the nature of the distinction and can easily be revised while elaborating the lexicon.

Finally, ordinary dictionaries contain a type of entry which only serves to refer the user to another entry constituted by the actual lemma. For instance, the entry *worse* in an English dictionary would do little more than refer to the lemma *bad*. This type of entry is unnecessary in a database because the search for words need not be limited to the field 'lemma' and can easily be extended to fields such as 'orthographic variants', 'phonological variants' or 'irregular inflection'. If the database is to be printed out in the form of a dictionary, non-lemmas can be generated from the items contained in these fields.

The above is the structure of a lexicon during elaboration. A published dictionary is a subset of the entries such that each printed entry contains a proper

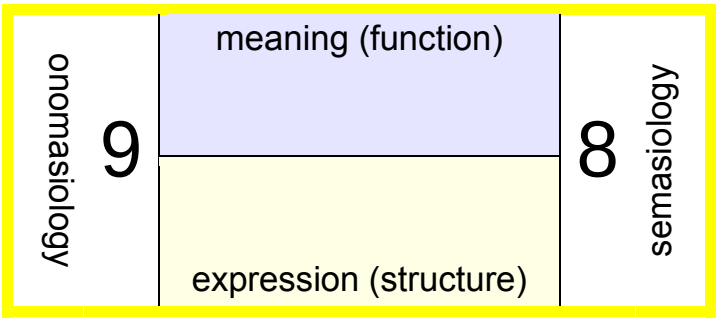
subset of the fields enumerated in T2. The entries are sorted according to a subset of the fields contained, typically by fields #1, 2 and 10.

5 Grammar[11]

In the case of the lexicon, the unit of description is obvious: it is the lexical entry. Its counterpart in the case of the grammar is not so obvious. A possible basic unit of grammatical description is the grammatical category, at all grammatical levels, from the morphological category via word class to the syntactic category of complex constructions.

If one proceeds in this way, one essentially produces a structural grammar. However, in reality the grammar of a language is its system of associating expression and content (as implied by the left-hand side of S2). Expression, i.e. formal grammatical structure, and content, i.e. grammatical meaning and function, form two frameworks which are mutually independent, but interrelated in multiple ways. Consequently, the grammatical description is situated in the intersection of a formal and a functional framework as shown in S4.

S4. *Onomasiological and semasiological perspective*



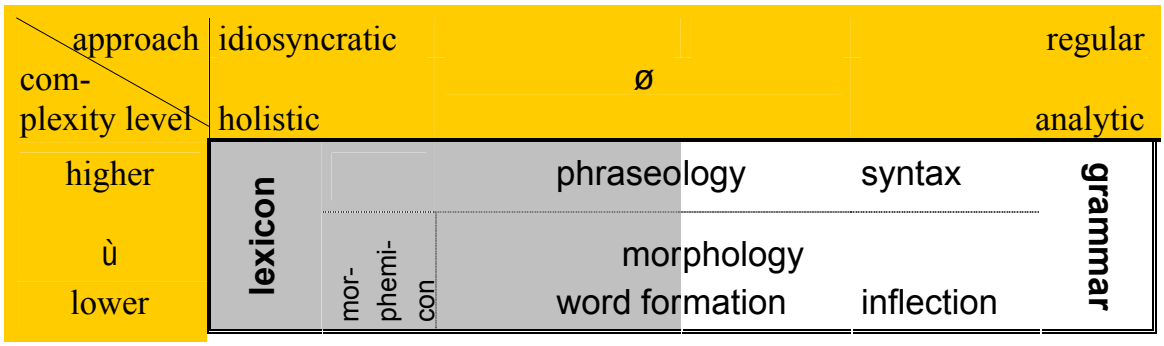
As has been well-known for a long time, S4 applies both to lexical and to grammatical description. The 'function' plane of a grammar corresponds to the subsection 'meaning' of T2; and the 'structure' plane of a grammar corresponds to the subsection 'structure' of T2. Moreover, the phonology and orthography (right-hand side of S2) that could prolong S4 at the bottom correspond to the subsection 'expression' of T2.

Both in the structural and in the functional grammar, there is a **taxonomy** and a **meronomy** of concepts. Thus, a record describing a word class such as 'adverb' contains a link of schema instantiation to the record describing the syntactic category 'adverbial' (cf. section 2.1). Again, a record describing the functional category 'progressive aspect' bears a relation of hyponymy to the

record describing 'imperfective aspect'. Similarly, just as a lexical stem may consist of two components (cf. field #11 of T2), so a syntactic category such as 'PrepP' bears meronomic relations to the categories 'Prep' and 'NP'. Again on the functional side, 'reference' and 'predication' bear a part-of relation to 'propositional act'. Thus, grammatical concepts bear the same kinds of relation to each other as lexical entries; and these may be implemented in similar ways.

As shown in S2 above, grammar and lexicon together constitute the semantic system of the language system. Concentrating now on the lower part of S4, we can say that these two components are intimately interrelated in terms of their structure, as shown in S5.

S5. *Lexicon and grammar*



The horizontal axis of S5 is the locus of the distinction between lexicon and grammar. The fact that S5 shows no clear dividing line between the two sides and that a number of components cross-cut the virtual dividing line, symbolizes the fact that lexicon and grammar form an integrated whole. The vertical axis represents the hierarchy of grammatical levels, which is in principle the same for lexicon and grammar, although much lexicology has ignored the levels above the word (i.e. phraseology).

The functional and structural categories introduced in the grammar are used in the lexicon to classify entries from the semantic and structural points of view. Each semantic category appearing in T2, field #19 is a link to a semantic category described in the grammar; and similarly, each structural category appearing in T2, fields #9–15 is a link to a structural category introduced in the grammar. In this way, lexical entries form the bottom of the grammatical taxonomy and meronomy, as indicated in T1. Technically, this means that the analyst working on the grammar may view all the members of a given category, and if he categorizes a lexical entry, the categories are provided by the grammar.

6 Text corpus

As said in section 1.1, the text corpus is the documentation proper. The corpus is composed of texts. One part of it are real running texts; the other part consists of a collection of example sentences or material elicited in the field, which is treated technically like a text. Each text is composed of units which are, in principle, sentences. For several practical reasons, they may be smaller syntactic units. Each such unit establishes a record in the database. The records are numbered consecutively.

Data are represented at various levels, from raw data over successively abstract representations up to annotations. Correspondingly, records are composed of fields which embody these representation levels and which are called 'tiers' in T3. Just as in T2, the set of fields represented in T3 is a maximum set, from which only a subset will be needed in any particular documentation. The last column of T3 refers to other components of the presentation of the language as introduced in S1 and S2. They are established on the basis of the type-token relationship between a unit of the description and a unit of the documentation, as shown in T1.

T3 gives an example of a record with the ID EMB 0253, which is the 253rd utterance recorded from the informant identified in field #14. Fields #2 – 10 constitute what are properly speaking tiers of representation of the data. Field #2 provides a link to an audio file. In this particular case, both this field and field #17 are empty because the utterance was only overheard. All of the data in the following tiers are, technically speaking, copies of the content of the corresponding fields of the lexicon.

The descriptors of field #11 serve the purpose of labeling a record for retrieval as an illustration of some phenomenon dealt with in the description. Ideally, these could be the converse of the example links referring from lexicon and grammar to the text corpus (cf. section 7). Field #12 clarifies the speech situation. For those subsets of the corpus which are literally texts, fields #13 and 14 will be filled in only once (in their "null record"[12]). The subsequent fields pertain to the working routine; field #18 is the counterpart of field #24 in T2.

[12] The ID of the first record of each text in the corpus contains the number 000. It gives information on the text as a whole.

T3. *Record of data in the corpus*

Level 1:	**Documentation**		
tier nº	tier name	example	related fields
1	record identity	EMB 0253	
2	acoustic/audiovisual record		
	representations:		
3	orthographic	Bik lúu'kech!	lex. 1
4.a	phonetic: segmental	[[ik ˈlúːket•]	lex. 5a
4.b	phonetic: prosodic	% ‾ _	lex. 5a
5	phonological: phonemic	/ bik lúːʃke… /	lex. 7
6.a	morphological: allomorphic	bik lúu'-k-ech	
6.b	morphological: morphemic	{ bik lúub-Vk-ech }	lex. 11a
7	interlinear morpheme gloss	PROHIB fall-SUBJ-ABS.2.SG	lex. 18
8	grammatical analysis	PTL V_{fin}	lex. 10, gr
9	native translation	¡Que no te caigas!	(lex. 17.b)
10	translation in background language	Don't fall!	(lex. 17.c)
11	descriptors	prohibitive, syncope	gr.
12	pragmatic comment	warning uttered on a pickup	soc. sit.
13	text genre	exclamation; instruction	
	field notes:		
14	speaker	Ernesto May Balam	data
15	recording date	01.08.89	

16	analyst	CL	
17	tape identity	-	
18	questions and problems	Why is syncope obligatory?	meth.
19	last modified	27.02.95	

7 Relations between lexicon, grammar and the other components

The relations between the lexicon and the other components are shown in the last column of T2. It is evident that the lexicon has interfaces with all the major components of the presentation of a language. Most important, perhaps, is the interface with the grammar, which is provided by the references to grammatical categories and constructions made in the fields of the subsection 'structure' of a lexical entry. As we saw in section 5, this reference is bi-directional.

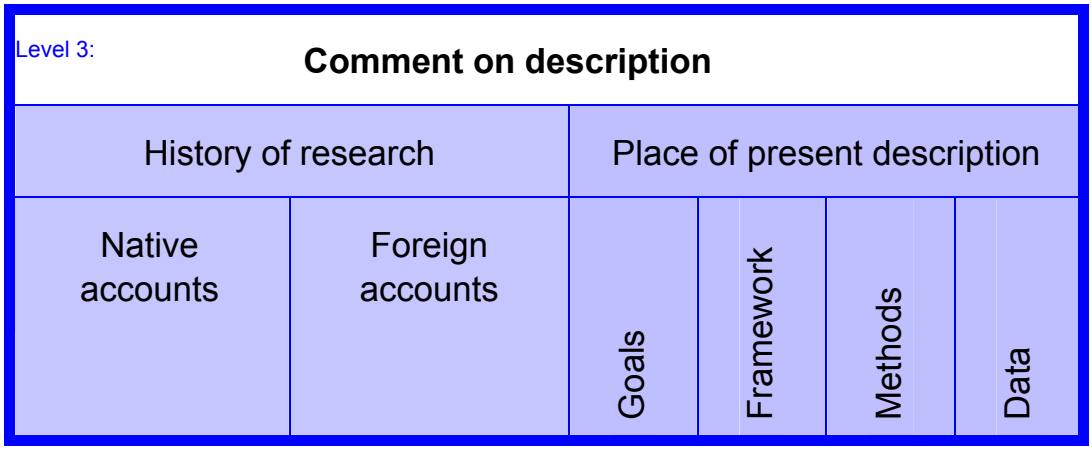
Furthermore, both the lexicon and the grammar make use of examples and of bibliographical references or, more generally, of sources of information. However, an example does not illustrate a lexical entry or a grammatical category as a whole, but only one aspect of them; and analogously for an information source. As said at the end of section 4.2, these two kinds of information therefore do not take the form of separate fields in the structure of the lexicon and the grammar and instead are inserted in diverse fields as specially formatted links to other components of the presentation of the language.

Each lexical and grammatical record has an associated set of **examples**. The examples are not given literally, but represented by record IDs of the text corpus. In other words, an example in the lexicon or the grammar is a link which points to the ID of a particular record of the text corpus (line #1 of T3). Conversely, most of the fields of T3 refer back to the lexicon by the simple fact that they contain tokens of the types embodied in a lexical entry. Thus, the relation between passages of the text corpus and lexical entries is a many-to-many relation. Analogously, the **information source** is a link to the bibliography. In the lexicon, it is indicated, in particular, for variants. The source of an example appears in the text corpus and need not be indicated in the passage where the example is used.

8 Comment on description

The activity of documenting and describing a particular language pursues a certain goal under certain conditions and in a certain historical, sociological and scientific context. The author of the activity owes the user of his product some reflection on these circumstances. S6 shows the subdivision of this component of the overall presentation of a language.

S6. *Comment on description*



First, a sketch of the **history of research** on the object language is provided. The subsection on native accounts is typically empty for languages which lack a written tradition; but there may be some folk linguistics. The **goals** of the documentation and description may be manifold. To render them explicit can be essential for the user in understanding what the author means with his account and in assessing it. The **theoretical framework** or model adhered to is identified, because it will be unfamiliar to future generations of scientists, let alone laymen. The **methods** applied in the collection, elicitation and analysis of the data are described so the user can assess the reliability of the account. Finally, the kinds and sources of **data**, including the informants with whom the linguist worked, are characterized. All of this constitutes necessary background information for the future use of the presentation of the language.

9 Conclusion

In a traditional linguistic description, which materializes in the form of a book, textual cross-references are a contingent phenomenon. In a comprehensive presentation of a language which takes the form of a database, they are essential and constitutive. They allow the analyst to structure his documentation and

description in a systematic fashion, and they allow the user of the presentation to see every linguistic phenomenon in the center of a network of relations. Thus, the components of the presentation are linked to each other in many ways, which are here briefly summarized:

Grammar and lexicon are related by the fields of the subsection 'structure' of a lexical entry. In the lexicon, such a field can only contain an item of the set defined in the grammar. Thus, the user can pass from a category of grammar to the words of the lexicon which belong to this category, and backwards.

The text corpus is related to all the parts of the description of the language (S2 and S3), in particular to both the lexicon and the grammar, by 'example' links which refer from a lexical entry or from a grammatical construction to a passage in the text corpus which contains the example.

The ethnographic situation of the language is related both to the lexicon and to the text corpus by links contained in the 'encyclopedic information' field of a lexical entry or in the 'pragmatic comment' field of a passage of text and which provide ethnographic background.

All components of the presentation are linked to the bibliography by links which provide a bibliographical reference.

References

Bohnemeyer, Jürgen & Lehmann, Christian & Verhoeven, Elisabeth 1994, "Situation of the language". AVG Arbeitspapier 6.

Comrie, Bernard & Croft, William & Lehmann, Christian & Zaefferer, Dietmar 1993, "A framework for descriptive grammars." Proceedings of the International Congress of Linguists 15(1992), vol. 1:159-170.

Langacker, Ronald W. 1987, Foundations of cognitive grammar. I: Theoretical prerequisites. Stanford, Calif.: Stanford University Press.

Lehmann, Christian 1982, "Directions for interlinear morphemic translations." Folia Linguistica 16:199-224.

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

Lehmann, Christian 2001, "Language documentation: a program". Bisang, Walter (ed.), Aspects of typology and universals. (Studia Typologica 1.) Berlin: Akademie Verlag; 83-97.

Lehmann, Christian 2002, "Documentation of grammar." Miyaoka, Osahito (ed.), Lectures on endangered languages: 3. From Kyoto conference 2001. Kyoto: Osaka Gakuin University (Endangered Languages of the Pacific Rim Publication Series, C003).

Seiler, Hansjakob 1969, "On the interrelation between text, translation, and grammar of an American Indian language." Linguistische Berichte 3:1-17.

Tsunoda, Tasaku 2001, "Role and ethics of researchers and method of documentation." Sakiyama, Osamu (ed.), Lectures on endangered languages: 2. From Kyoto conference 2000. Kyoto: Osaka Gakuin University (Endangered Languages of the Pacific Rim Publication Series, C002); 261-268.

Wimbish, John S. & Davis, Daniel W. 1992, Shoebox. Integrated data management and analysis for the field linguist. [Austin, Tx.]: Summer Institute of Linguistics.