CLIPP

Christiani Lehmanni inedita, publicanda, publicata

titulus	
	Documentation of grammar
huius textus situs retis mundialis	
dies manuscripti postremum modificati	
	18.09.2002
occasio orationis habitae	
	Conference on Language Endangerment, Kyoto, 30.11.– 02.12.2001
volumen publicationem continens	
	Sakiyama, Osamu & Endo, Fubito & Watanabe, Honore & Sasama, Fumiko (eds.), *Lectures on endangered languages: 4. From Kyoto Conference 2001*. Osaka: Osaka Gakuin University (Endangered Languages of the Pacific Rim Publication Series, C-004)
annus publicationis	
	2004
paginae	
	61-74

Documentation of grammar

Christian Lehmann

University of Erfurt

Abstract

Documentation of the grammar of a language is a constitutive part of a comprehensive presentation of a language. The grammatical system is approached from two opposite perspectives, the onomasiological (or functional) and the semasiological (or structural) perspective. They are based on language-independent frameworks and thus allow an unbiased account of the language system.

At the same time, the two perspectives are associated with deductive and inductive methodology, respectively. The former operationalizes the concepts and operations of functional domains in the form of questionnaires and translation sentences. The latter applies the methods of structural analysis to a text corpus.

1   Documentation and description of a language

Since language endangerment has been enjoying increased attention both in linguistics and in the general public, it has become customary to speak of the 'documentation of a language' as an activity by which linguists react to the situation (cf. Lehmann 2001, Tsunoda 2001). However, as the expression has gotten en vogue, it has become decreasingly clear what it means. One often has the impression that 'language documentation' is a fashionable substitute for what used to be called 'language description', and we use the fashionable expression because it sells better. However, it is important to make a distinction between documentation and description of a language, not only for the sake of terminological clarity, but also for methodological and practical reasons (cf. Himmelmann 1998). I will first draw the distinction and then come back to its practical import.

The distinction between documentation and description of a language pertains to the methodological level: A **documentation** of a language is a presentation of linguistic data in the form of a text corpus representing communicative events in a speech community. Its object is, thus, speech (*parole*), not language (*langue*). It contains raw data (i.e. video or audio recordings) and linguistic representations of these at various levels of abstractness, possibly with annotations (cf. Lehmann 1983, Bak-
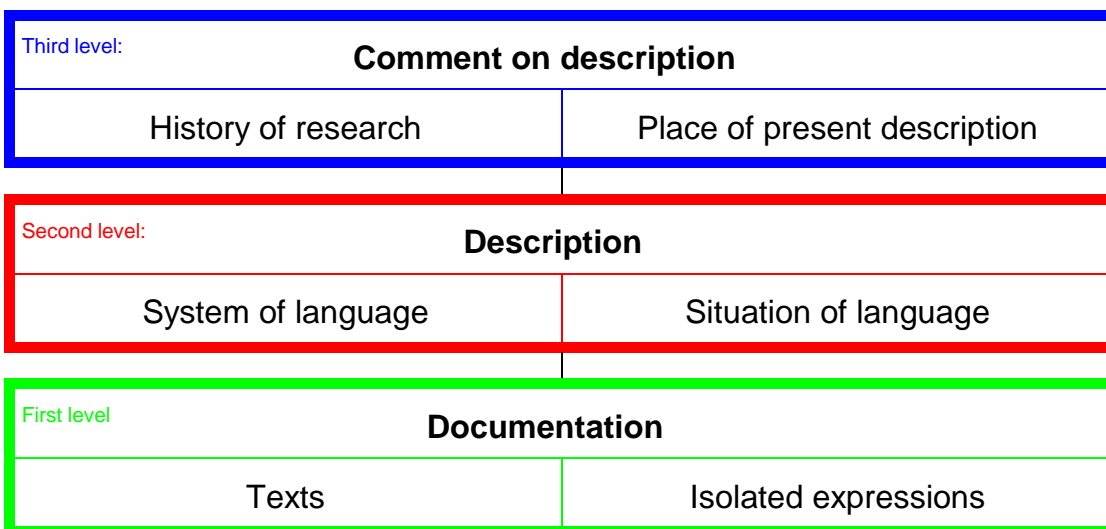
ker et al. 1994, sections 3f). Its structure is dictated by the sequence and structure of texts that it consists of.

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

A description is, thus, at a meta-level with respect to a documentation. 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.

Above the two levels of documentation and description of the language, the presentation comprises a third level which again bears a meta-relation to the second (and first) level. These are the **comments on the description**, by which the researcher accounts for his research, its conditions, its purpose and its conduct. The three levels of a comprehensive presentation of a language are displayed in S1.

S1. *Presentation of a language*

Third level: **Comment on description**	
History of research	Place of present description

Second level: **Description**	
System of language	Situation of language

First level **Documentation**	
Texts	Isolated expressions

I am coming back to the practical necessity of distinguishing between documentation and description. Working linguists' experience with the elaboration of edited, annotated and translated text corpora, of grammars and dictionaries teaches us that the comprehensive presentation of a language can easily occupy a linguist's lifetime. A language is more complex than a linguist. However, in the case of an endangered language, we do not have the lifetime of a linguist at our disposal. The language can die

out within a few years. A responsible planning of the work, with a feasible sequencing of the necessary steps, therefore becomes of the utmost importance.

Let us assume the concepts of documentation and description as defined above, and let us call a sufficient documentation one on whose basis one can elaborate a description of the language. Now, it is possible to come up with a sufficient documentation of a language within a few years. If the language then becomes extinct, it will still be possible to elaborate its description at leisure. If, on the other hand, we concentrate our forces on describing the language, and it dies out after a few years, then we do not have a full description, and even worse, it is not possible to generate a documentation on the basis of an existent description. This is the practical reason why, if we are concerned with the documentation of endangered languages, we have to mean 'documentation' in the narrow sense defined here.

In this conceptual framework, the expression 'documentation of grammar' so far has no interpretation, since a grammar is part of the description, not of the documentation of a language. I will give this expression the following meaning: By **documentation of grammar** is meant a part of the documentation of a language which generates data that are necessary and sufficient for the elaboration of the grammar of the language. The documentation contains, so to speak, the **interface for the grammar**.[1]

2   Theoretical prerequisites for documenting grammar

2.1 The position of grammar

In order to understand the methodology of grammar documentation, we have to review briefly the position of the grammar in the language system. As shown in S2, the language system has two main parts, the semantic system and the expression system. Grammar and lexicon together form the semantic system. Thus, 'grammar' will here be used in a narrow sense that excludes phonology and orthography.

---

[1]   Cf. Himmelmann 1993 for a similar conception.

S2. *System of language*

Level 2: Description	**System of language**		
Semantic system		**Expression systems**	
Lexicon	Grammar	Primary: Phonology	Secondary: Writing

The two semantic subsystems of the language system have the task of mapping meaning onto sound, of associating function with structure. At first blush, there appear to be infinitely many ways a language can do this. The famous passage by Martin Joos quoted in Q1 is a vivid expression of the apory felt here.

> Q1. *Martin Joos in his comment on Bloch 1941 in Joos (ed.) 1957:96*
> Trubetzkoy's phonology tried to explain everything from articulatory acoustics and a minimum set of phonological laws taken as essentially valid for all languages alike, flatly contradicting the American (Boas) tradition that languages could differ from each other without limit and in unpredictable ways, and offering too much of a phonological **explanation** where a sober taxonomy would serve as well.

On the other hand, the linguist has the task to present a language in such a way that it becomes understandable and comparable with other languages. Thus, we can invoke Immanuel Kant in Q2 against Martin Joos:

> Q2. *Categorical imperative of language description*
> Describe your language in such a way that the maxim of your description could serve, at the same time, as the maxim of the description of any other language.

In order to reconcile the tension between universality and diversity, we need an approach to linguistic description which presupposes what is universal in language and on this basis allows us to bring out the peculiarity of the particular language.

2.2 Functional and structural grammar

At the level of grammar, the categorical imperative of language description can be complied with in the following way:[2] The **association of function with structure** is achieved at the level of the individual language. The bases of linguistic functions and of linguistic structures themselves, however, are universal. That is, there is a universal set of functions fulfilled by human cognitive and communicative acts. And there is a universal set of expression structures that serve in human semiosis. In a given language, each function is associated with one or more structural devices; and each structural device serves one or more functions. There is thus a many-to-many mapping between functions and structures, as illustrated in S4 below.

The association of function with structure is performed by the speaker and by the hearer in converse directions. The speaker starts with a concept, a cognitive or communicative function that he wants to fulfill and looks for the expression or the structural devices that the language offers him for this purpose. If he is a speaker of Japanese expressing himself in English, he will use a Japanese-English dictionary to find the expressive means that suit his purposes. The hearer is confronted with expressions and structural means and looks for the meanings or functions that they fulfill in the language. If he is a speaker of Japanese reading in English, he will use an English-Japanese dictionary to find the concepts and functions associated with the English expressions.

Just as a dictionary consists of the two volumes alluded to, a grammar may be arranged according to the same two complementary principles. It is either based on the cognitive and communicative functions, being organized and subdivided according to their intrinsic logic, and looks for the structural devices that the language uses to fulfill them. This is called the **onomasiological** or **functional** perspective. Or else it is based on the structural devices, being organized and subdivided according to their intrinsic logic, and looks for the functions that the language fulfills by them. This is called the **semasiological** or **structural** perspective. The two perspectives are displayed in S3 (cf. Lehmann 1980, 1989 and Comrie et al. 1993).

---

[2]   See Lehmann 1980 for the following.

S3. *Onomasiological and semasiological perspective*
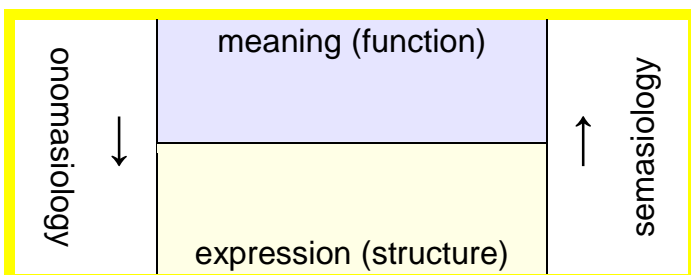
Turn S3 clockwise by 90° to derive the example from Latin grammar illustrated in S4. In the onomasiological perspective, we may start with a concept such as the possessive relation and look for its materializations in Latin. We find, among other things, the structural devices enumerated on the right hand side of S4. In the semasiological perspective, we may start with a structural category such as the genitive and look for its functions in Latin. We find, among other things, the functions enumerated on the left hand side of S4. The possessive genitive (second row of S4) emerges as a specific association of a function with a structure which has its place in a complex network that can only be untangled by separating the two perspectives.

S4. *Onomasiology and semasiology of the Latin possessive genitive*

conceptual relations	mapping	structural devices
partitive (X part of Y)		Y-DATIVE *est* X-NOM
possessive (Y possesses X)		Y-GENITIVE X
explicative (Y explicates X)		Y-DATIVE verb X
material (Y is material of X)		Y-NOM *habet* X-ACC
onomasiology	→ ←	**semasiology**

The example shows that the two perspectives are completely independent and complementary of each other. A comprehensive grammar has to take both perspectives.

A semasiological or **structural grammar** is organized by a purely structural logic, summarized and illustrated in T1. Its basis is formed by grammatical units such as morpheme, stem, word-form, phrase and clause (some of which are not universal). At each level constituted by such a unit, there is a set of language-specific categories, e.g. word classes. Thirdly, a unit of each of these categories may be internally complex, forming a construction with internal relations between members that have a certain distribution and where the operator has a certain structural representation vis-à-

vis the operand. Last, the operator expands into a paradigm of units which occasion variation on the operand or exhibit variation occasioned by the operand.

T1. *Structural grammar*

nº	perspective	comment	examples
1	Grammatical units	units constituting grammatical levels	sub-morpheme, morpheme, stem, word, phrase, clause, sentence
2	Grammatical categories	distributional classes of units at each level.	verb stem, finite verb, verb phrase, verbal clause, cleft-sentence
3	Syntagmatic structure		
3.1	Relations in constructions	sociation, dependency	apposition, direct object
3.2	Distribution	position, obligatoriness of components	optional prenominal position
3.3	Operators	grammatical formative forming a (binary) construction with the operand	sign with segmental representation, affixation, modification
4	Paradigmatic structure		
4.1	Members of the paradigm	values of a grammatical category	case-number suffixes
4.2	Variation in members of the paradigm	allomorphy in the operator	number allomorphy conditioned by gender
4.3	Variation in the operand	allomorphy of stem	noun stem allomorphy conditioned by number

An onomasiological or **functional grammar** is organized by a purely conceptual, operational logic, summarized and illustrated in T2. At the most general level, cognitive and communicative functions are organized in cognitive and communicative domains,[3] such as possession, quantification etc. At subordinate levels, these comprise particular concepts such as time, agentivity, inalienability etc., concepts that are manifested at the grammatical level in many languages. At the same time, there are operations to generate and modify such concepts, such as predication, determination, abstraction etc.

---

[3] comparable to the functional dimensions in the sense of Hansjakob Seiler (2000)

T2. *Functional grammar*

Functional domain	Main areas
Apprehension and nomination	Categorization systems, conceptual types
Concept modification	Attribution, apposition
Quantification	Quantification in reference, quantification in predication
Reference	Determination (incl. deixis), reference tracking
Possession	Possession in reference, possessive predication, possession and participation
Space construction	Orientation points, local relations, spatial regions, spatial and gestalt properties of objects
Predication	Existence/placement, characterization, participation
Design of situations	Aspectuality, classes of situations, modification of situations (aktionsarten)
Temporal orientation	Absolute time, temporal relation
Illocution and modality	Declaration, question, exclamation, request and command, hortatory, monitory, obligation, volition, possibility, evidential, shading
Contrasting	Negation, comparison, gradation
Nexion	Interpropositional relations (semantic coordination, semantic subordination)
Articulation of discourse	Discourse structure, Functional Sentence Perspective

T2 forms the top of S3, while T1 forms its bottom. The general bases of the two systems as exhibited in T1 and T2 are universal. However, to the extent that they are spelled out in detail (moving vertically in S3), language-independent concepts and operations and language-independent structural devices stepwise acquire language-specific properties, such as, e.g., the categories mentioned in S4. Thus, each of the two systems may be viewed as a hierarchy whose top is universal, but whose bottom is language-specific. Each individual sign of a language, be it a word, a morpheme or a construction, is thus situated at the bottom of both hierarchies and, therefore, at the intersection of functional and structural categorizations.

3   Methods for the documentation of grammar

In linguistic fieldwork, documentation and description, just as everywhere in linguistics, **inductive and deductive methodology** have to complement each other. They can be systematically associated with the two perspectives of S3 as shown in T3.

T3. *Perspective and method in grammatical analysis*

basis	viewpoint	perspective	method
forms and structures	hearer	semasiological/structural	inductive
cognitive and communicative functions	speaker	onomasiological/functional	deductive

3.1 The inductive method

The inductive method in language documentation takes the viewpoint of the hearer. For this method, the representativeness of the data is of prime importance. The first step is therefore the recording of a corpus of utterances and texts that represent the various speech situations and genres which are traditional in the society (cf. Lehmann 2001).[4] The texts are then segmented, the units are identified and classified by the application of methods of substitution and permutation. Major class words are inflected through their various morphological categories in order to establish morphological paradigms, and transformations are applied to sentences in order to establish syntactic paradigms. In this way, the levels, categories, syntagmatic relations and paradigms of structural grammar in the sense of T3 are established, much as structuralist schools of the first half of the 20th century from L. Bloomfield to Ch. Hockett used to teach.

It is not necessary to illustrate this well-known inductive methodology here. It goes without mentioning that it does not have the status of a discovery procedure. It is merely a heuristics, and an incomplete one at that, because it has to be complemented by deductive methodology. The results produced by this methodology are not yet a grammar; recall that we are talking not about grammar, but about the documentation of grammar. Rather, they constitute the raw material on whose basis a structural grammar can be elaborated.

---

[4]   The corpus may also contain isolated utterances and elicited sentences, as suggested in Himmelmann 1993.

3.2 The deductive method

The deductive method takes the viewpoint of the speaker. There are various ways of applying it in language documentation. One way is to convert the functional concepts and operations into a questionnaire. An early example of this is the "Lingua Descriptive Studies Questionnaire" by Comrie & Smith (1977). Ideally, such a questionnaire is subdivided according to such functional domains as enumerated in T2. Each domain is spelled out down to the level of typological grammatical categories. For instance, in the domain of Design of Situations, subdomain of Aspectuality, it may be asked whether the language makes a grammatical distinction between such dynamic situations which are bounded and such which are unbounded. If there is such an aspectual category, it may then be named 'perfective vs. imperfective' or 'plain vs. progressive' or 'completive vs. incompletive', corresponding to diverse shades in function and descriptive traditions.

Questions such as those in the "Lingua Descriptive Studies Questionnaire" are yet at a metalinguistic level, so that they can only be understood and answered by a trained linguist. A further step in guiding the inexperienced fieldworker may be taken by operationalizing such questions in the form of test frames in which expressions of the object language have to be inserted. For instance, terminative verbs can be distinguished from durative verbs by a test frame such as S5.

S5. *Test frame for terminative vs. durative aktionsart*

$$\text{Peter Ved (the X) } \left\{ \begin{array}{l} \text{(with)in} \\ \text{for} \end{array} \right\} \text{ three hours/seconds.}$$

S5 is to be translated into the target language. Then the verb whose aktionsart is to be tested is inserted in place of V, and X is replaced with an appropriate noun if V is transitive. If the first version is good and the second version bad, the verb is terminative (e.g. Engl. *burn*); if the result is the reverse, the verb is durative (e.g. *walk*). See Lehmann 1993 for more such test frames and for the problems associated with their use. As an aside, it is interesting to observe that this particular method figured already in structural linguistics. As a matter of fact, it presupposes some prior hypothesis from which it follows that expressions of a certain category will fit in the test frame, while others will be incompatible. This is actually possible only in the deductive method.

The final step in enabling the layman to provide the material sought for in the onomasiological approach consists in the elaboration of model sentences which are to be translated. An early example of this method is the series *Archivo de Lenguas Indígenas de México* launched by Jorge A. Suárez and now edited by Yolanda Lastra

(1974ff). Here, the documentation of a language consists of the translation of a set of several hundred standardized sentences into the target language. The sentences are chosen in such a way as to maximize chances that their translations will exhibit the central grammatical categories and vocabulary of the language.

This method has been further refined by Östen Dahl, for instance in Dahl (ed.) 2000, appendices. Here, characteristic little stories or situations are constructed in which the sentence to be translated is embedded. The context is configured in such a way as to force the association of that sentence with the cognitive category which is at stake and whose expression in the target language is to be tested. In the English original version of the questionnaire, the category would appear in its English grammatical manifestation, but that is suppressed by presenting the host word as a mere vocable, without any grammatical categories and, in particular, without any hint to the grammatical category being tested.[5] Q3 is a typical example from such a questionnaire.

Q3. *Progressive, imminential meaning (Dahl (ed.) 2000:813)*

S56. Hurry up! The train LEAVE
S57. The old man DIE, but finally they found the right medicine.

In the example of Q3, the imminential progressive is presupposed as a functional category, and it is asked which structural category the target language uses to express it.

Another set of methods within an overall onomasiological approach involves the elicitation of linguistic behavior by non-linguistic stimuli. The Max Planck Institute for Psycholinguistics at Nijmegen has been developing, over the years, a sizable set of tools, kits and experiments to be employed for this purpose in diverse cognitive and communicative fields. One type of methods involves the representation of little scenes with puppets or by silent movies, which are then to be described or retold by the native subjects. There may also be communicative problems to be solved, such as the task to orient a fellow in space or to instruct him to mount a device. All of these methods presuppose a certain functional domain and a set of cognitive or communicative operations in it. The setup of the experiment is designed, and the task is defined in such a way as to guarantee that the linguistic solution to the task will make use of the grammatical devices that the object language possesses in that area.

The advantage of this type of methods against those methods which involve translation is that it excludes interference from other languages. Its disadvantage is, of course, that it is relatively costly in terms of time and money.

---

[5] This is, of course, done in order to minimize interference from the metalanguage used.

Again, it goes without saying that none of these methods is a discovery procedure. They have the status of a heuristic which provides material for the documentation of the grammar of the language. The material has to be interpreted and systematized, and the deductive methodology has to be complemented by the inductive methods mentioned before.

3.3 Grammar and lexicon

We saw in S2 that the grammar of a language is most intimately connected with the lexicon. The lexicon contains not only words, but also stems and bound morphemes. Each entry forms a union of expression and meaning. Moreover, each entry has various grammatical properties, such as its distribution class, its inflection class or its construction frame. If the lexicon is implemented as a database, as it will be until its eventual publication in the form of a dictionary, then each entry may be accessed by the onomasiological or the semasiological approach.

In elaborating the description of a language, it is wise to associate as many diverse kinds of information as possible with the entries of the lexicon. A database program allows the user to limit the set of permissible entries of a certain field to a closed set of items. For instance, the field 'part of speech' of lexical entries may only contain categories such as 'inalienable noun', 'ditransitive verb', 'clitic personal pronoun' and a limited number of others. The field is, in this sense, standardized. The same goes for a considerable number of fields in a lexical record which have it in common that they classify the entries by diverse criteria. In the elaboration of the lexicon, the set of items actually needed in each such classificatory field gradually emerges in an inductive way. At appropriate stages in the process, the analyst reviews the categories he has used up to now, switches to the deductive perspective and transforms the set into a coherent classification. This alternation of inductive and deductive method in the elaboration of the system embodied by the lexicon proceeds in the form of a spiral until the analyst quits.

At this point, the lexicon contains the part-of-speech system, the inflectional categories, the semantic categories, the morphological categories and their members, the word-formation processes, and many other systematic aspects of the description will be embodied in the lexicon. The subsystems mentioned, however, at the same time form the core of the grammatical description of the language. Thus, a way of approaching the grammatical description of a language is to elaborate its lexicon; one then gets a large portion of the grammar for free.

Naturally, this does not argue for completing the lexicon before the grammar in a language description. It rather goes to show that if, in the elaboration of a lexicon, the

analyst makes clever use of a database program, its in-built properties help him to achieve the degree of systematicity and consistency necessary for a grammar.

4   Conclusion

The most important points may be summarized as follows:
- The documentation of a language must be distinguished from its description. The grammar is essentially part of the description and may only be prepared by the documentation.
- A grammar is constructed in such a way as to allow an onomasiological and a semasiological perspective on its content. A language documentation that provides the interface for a grammar therefore has to apply both inductive and deductive methods.
- Furthermore, a broad foundation and a major portion of the work required for the grammar is achieved if entries of the lexical database have a rich field structure which includes diverse kinds of classificatory information.

References

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.

Comrie, Bernard & Smith, Norval 1977, "Lingua Descriptive Studies: Questionnaire." *Lingua* 42/1.

Dahl, Östen (ed.) 2000, *Tense and aspect in the languages of Europe.* Berlin & New York: Mouton de Gruyter (Empirical Approaches to Language Typology, EUROTYP 20-6).

Himmelmann, Nikolaus P. 1993, "Material ambitions." *Languages of the World* 7(2):66-68.

Himmelmann, Nikolaus 1998, "Documentary and descriptive linguistics." *Linguistics* 36:161-195.

Joos, Martin (ed.) 1957, *Readings in linguistics. The development of descriptive linguistics in America since 1925.* New York: American Council of Learned Societies.

Lastra, Yolanda (ed.) 1974ff, *Archivo de lenguas indígenas de México.* México, DF: Colegio de México.

Lehmann, Christian 1980, "Aufbau einer Grammatik zwischen Sprachtypologie und Universalistik." Brettschneider, Gunter & Lehmann, Christian (eds.), *Wege zur*

Universalienforschung. Sprachwissenschaftliche Beiträge zum 60. Geburtstag von Hansjakob Seiler. Tübingen: G. Narr (TBL, 145); 29-37.

Lehmann, Christian 1983, "Directions for interlinear morphemic translations." *Folia Linguistica* 16, 1982:193-224.

Lehmann, Christian 1989, "Language description and general comparative grammar." Graustein, Gottfried & Leitner, Gerhard (eds.), *Reference grammars and modern linguistic theory.* Tübingen: M. Niemeyer (Linguistische Arbeiten, 226); 133-162.

Lehmann, Christian 1993, "Predicate classes in Yucatec Maya." *Función* 13/14:195-272.

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

Lehmann, Christian & Bakker, Dik & Dahl, Östen & Siewierska, Anna 1994, *EUROTYP Guidelines.* Strasbourg: Fondation Européenne de la Science (EUROTYP Working Papers).

Seiler, Hansjakob 2000, *Language universals research. A synthesis.* Tübingen: G. Narr (Language Universals Series, 8).

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.