Hans-Heinrich Lieb
Sebastian Drude

Advanced Glossing:
A Language Documentation Format

(Working Paper, November 2000)

Table of Contents

1 Introduction

1.1 Requirements for a language documentation format

We consider the following conditions (which may not be independent) as minimal requirements that any language documentation format must meet if the documentation is to be suitable for the purposes of linguistics:

(1) It must be possible to write a grammar of the language or variety being documented given a sufficiently large number of texts completely documented within that format.

(2) The language used as the documentation language must be clearly interpretable; in particular, it must be possible to clearly distinguish between phonetic, phonological (phonemic), morphological, syntactic, and semantic information in the documented text.

(3) The documentation format must allow both for partial documentation of a text and for the gradual, systematic filling in of gaps during the documentation process.

(4) The documentation format must be such that information gathering for the complete documentation of a text (which may not be possible in all cases) can, in principle, be achieved under field conditions.

We wish to emphasize the character of these conditions: they are (i) minimal and (ii) conditions for the purposes of linguistics. Even for these purposes, additional conditions easily come to mind, and the above requirements do not yet cover the conditions imposed on language documentation for purposes outside linguistics. It may be argued, though, that the above requirements must also be imposed for many aims pursued in other fields.

1.2 Typological Glossing and Advanced Glossing

It should be safe to say that so far no systematic attempt has been made at developing a language documentation format that meets Requirements (1) to (4). In particular, interlinear morphemic translation as widely used in typological studies and first systematized by Lehmann (1982) — henceforth called "Typological Glossing" (TGl) — does not yet meet the fundamental Requirement (1), and may not meet (2). Indeed, Lehmann's own caveats have not been sufficiently heeded by those who subsequently adopted his proposals:

(i) Lehmann's proposals are restricted to 'interlinear *morphemic* translations', which by definition are restricted to morphology and whose relation to phonetics, phonology, syntax, and semantics is, therefore, indirect at best.

(ii) The tentative nature of Lehmann's proposals was largely forgotten: The reader is explicitly asked by Lehmann to "regard the following proposals as a preliminary version of something which will certainly be greatly modified before it can be called anything like final" (1982: 200).

On the other hand, TGl has doubtless proved useful in a typological context due to features that may also be relevant in a context of documentation (the two contexts must still be

carefully kept apart). Generally, TGl has proved itself in the highlighting of morphological or morphology-related features of sentences or their parts whenever other linguistic information on a sentence is either irrelevant for the purposes on hand or easily retrieved from context. It is therefore advisable — quite independently of DOBES requirements — to identify TGl features that should be retained for language documentation formats, and integrate them into proposals for such formats. Our own proposal will indeed integrate these features (without first isolating them individually). It is therefore a proposal for a glossing format, and will be called "Advanced Glossing" (AGl).

1.3 Terminological remarks

Henceforth, we will keep to the following terminological conventions.

We speak both of the *documentation* of a language or language variety and the *documentation* of a text. In either case "documentation" is to include all features the researcher is interested in, not only linguistic ones. The expression "text" will be used only for comprehensive entities that may constitute genres. Part of the documentation of a text may be the *glossing* of its sentences, either from a syntactic or a morphological point of view (*syntactic* glossing vs. *morphological* glossing). The morphological glossing of a sentence is achieved through the (morphological) glossing of its words. We will speak of *a glossing* or *glossings* to refer to individual descriptions of sentences or words in a glossing format. We will assume that each glossing consists of a table: a *glossing table*, and a comment on the table — a *Comment*, for short.

Relatively little will be said on the *Comment* of a glossing. It will become apparent, though, that only two parts need be assumed for structuring the Comment. The *first part* of the Comment is simply a list of entries that each consist of a name of one cell or one line in the glossing table (see below, Sec. 3.1, for the notion of cell) and a part that gives the relevant information. Due to the cell structure of a glossing table, links can be established between individual cells or lines of the table and individual entries in the list that is the first part of the Comment. The *second part* of the Comment should provide information not specific to cells of the table such as information on etymology (in the case of a morphological table), dialect, register etc.

Our proposals will be for a glossing format as one component of a general documentation format; other components will not be considered. (See Appendix 3 for an overview.)

1.4 Methodological remark

A documentation format is, or incorporates, also a format for *description*. Description formats must be carefully distinguished from research methodologies. Not infrequently, description formats are criticized on the basis of a tacit assumption on method: the assumption that the format is to be applied in an actual research situation by schematically following its outline. This may well be a rotten method. How the researcher proceeds is not defined by the format, just as little as the degree of completeness he or she strives for. While a description format must be methodologically sound, it should not be taken as a recipe for organizing actual research.

1.5 Presentation of Advanced Glossing

In this Working Paper, AGl will not be presented in full generality (an impression of the overall structure may be obtained from Appendix 3). Rather, we will use examples for explaining all important AGl features. No attempt will be made to explicitly compare AGl and TGl; the relationship should be obvious from the examples. (These will be drawn from an exotic language — currently in danger of being replaced by Pidgin English — German.)

The following Section 2 essentially consists of two glossing tables, one for syntax and one for morphology. The documentation situation is fictitious, that is, we assume certain raw data as given; they represent the starting-point for obtaining the glossings. The raw data consist of a taped speech event, preferably supplemented by the results of running the event through an automatic speech analysis program. The (elliptic) sentence realized by the speech event is deliberately simple, so as to avoid irrelevant complications in the syntactic glossing. The example chosen for morphological glossing is, however, fairly complex; here, it is triviality that had to be avoided. Reasons for separating syntactic and morphological glossings will be given.

2 Advanced Glossing

2.1 Syntactic glossings: Table 1

(See Appendix 1. [The Appendix constitutes a separate attachment.])

2.2 Notations used in Table 1

Phonetic symbols: A narrow IPA-transcription has been used; vowel length is indicated by symbol doubling; syllable boundaries are indicated by a dot to the right of a sound symbol or, in the case of linked syllables, beneath or above a sound symbol.

Phonemic and archiphonemic symbols: self-understood for a phonology of Modern Standard German. Phonological intonation structure is coded by the usual word-stress symbols. Hyphens in a phonemic transcription indicate the beginning of a syllable and are used to isolate preceding consonants that are extrasyllabic.

Pitch symbols: "L" for low pitch, "H" for high pitch, "M" for mid pitch, subscripted "f" for "falling", subscripted "r" for "rising".

Category symbols: "Unm_G" for "Unmarked for gender", "Str" for "Strong (adjective or pronoun form)", "Unm_C" for "Unmarked for case", "Wk" for "Weak (adjective or pronoun form)", "Unm_D" for "Unmarked for definiteness", "Nf" for "Noun form", "NGr" for "Noun group". The remaining symbols are self-explanatory. (Our category symbols are, of course, subject to standardization.)

Relation symbols: "mod" for "modifier".

Notation for concepts as lexical meanings: An English word that is suggestive of conceptual content is used between single quotation marks.

Language for sentence meaning paraphrases: same as for concepts.

Remark. If necessary, individual symbols in a glossing table may be defined in the Comment of the glossing.

2.3 Morphological Glossings: Table 2

(See Appendix 2. [The Appendix constitutes a separate attachment.])

2.4 Notations used in Table 2

Phonemic and archiphonemic symbols: same as in Table 1.

Pitch symbols: Same as in Table 1 except for "H,L", used in the underlying phonological theory to indicate secondary word stress. (Occurrence of H,L is, however, not per se sufficient for secondary word stress.)

Category symbols: The naming of individual categories in Table 2 is schematic because we do not here wish to go into the details of German morphological classes. Subscripts after an expression such as "Pref" (for "Prefix") simply indicate that a certain subclass of the category in question (here, the category Prefix) would have to be named. Other than that, the expressions used should be self-explanatory. "St" is short for "Stem"; the slash may be read as "is transformed into"; "Stf", "StGr" and "Af" are short for, respectively, "Stem form", "Stem Group" and "Affix form". — Category symbols taken over from Table 1 are interpreted as in Table 1.

Relation symbols: "m-mod" for "morphological modifier"; "m-qual" for "morphological qualifier".

Notation for concepts as lexical meanings: as in Table 1.

Notation for traditional grammatical meanings in derivation: Suggestive English words ("not", "suitable-for") are used without any further marking.

3 General comments

3.1 Shared features of the syntactic and morphological glossing tables

In agreement with Requirement (2), morphological and syntactic information is represented separately but in an interconnected way. The phonological words in the syntactic glossing of Table 1 are numbered, and this numbering should already be sufficient to associate morphological glossings, if available, with individual words.

Both the syntactic and the morphological glossing tables are maximal, that is, they are, in principle, examples of *complete* glossing tables. Due to the way the glossing format arranges relevant information, incompleteness of various kinds can also be covered, see below, Sec. 3.2.

Both the syntactic and the morphological glossing tables each consist of thirteen lines, here identified by Roman numerals that serve as names of the lines. Our choice of Roman numerals is ad hoc; eventually, names suggestive of the character of each line (see Secs 4 and 5) may be preferable and may be supplied in, for example, a Shoebox style.

Although the Arabic numerals in Line I are no metasymbols, they can be used as such: In both the syntactic and the morphological glossing tables an Arabic numeral in Line I together with one of the Roman numerals in Lines II to X identifies a cell. The information that may be represented in each one of these cells is cell-specific, as will be explained in subsequent Sections.

3.2 Incomplete glossings

The format allows for kinds of incomplete documentation in arbitrary combinations. For example, we may have information pertinent to cells II.1, VI.4 and IX.4. In this case, the relevant cells for the sentence in question may be filled whereas all other cells remain empty. Either the same or a different researcher may eventually be able to fill some or all empty cells. However, even if gaps remain, we know not only what the information is that we have got, we also know exactly what is missing, and this may direct additional research, even at a later time or by a different researcher. Depending on the type of available raw data, there may be systematic incompleteness insofar as entire lines of a table may remain empty. In some cases, due to more limited linguistic aims, certain lines or cells may be deliberately left empty, either completely or partly, because the information they contain is irrelevant to the aims.

Cells that are ‘empty’ due to lack of information should be marked as such, e.g., by means of a question mark.

Our examples of glossings are incomplete also by being restricted to glossing tables. As mentioned above, for each table there is a Comment. Uncertainty concerning cells of a glossing table may also be made explicit in the Comment.

3.3 Nature of glossings

It should be emphasized that all glossings, whatever the format or type, are bundles of hypotheses and do not register god-given truths. This holds even of a seemingly innocuous part of a glossing table such as the phonetic transcription of the raw data: in writing down or typing a phonetic symbol such as the letter “p”, the researcher, on the basis of a sound impression, formulates a hypothesis such as, a certain part of the speech event was caused by a complex articulatory movement involving closing of both lips without vocal cord vibrations, a hypothesis that may well be wrong. A glossing format should allow not only for the closing of gaps, but quite generally, for corrections.

4 The syntactic glossing table (1): Representing basic information (Lines I to IX, XII, and XIII)

4.1 Line I: Number and order of phonological words

The Arabic numerals indicate the number and order of the phonological words whose orthographic names are given in Line VI. The numbering embodies the hypothesis that there is a fixed number of phonological words that occur, in this case, 3. Insofar, the numbers are not part of the metalanguage. The hypothesis may well be wrong, and renumbering may be required. For this reason, the numbers should be assigned and changed automatically. Any doubt concerning the division into phonological words may be formulated in a part of the Comment that refers to Line I.

A glossing table may have to be split up into several parts when we are dealing with a longer sentence. Each part conforms to the general format, but the numbering in the various Lines I is consecutive.

4.2 Lines II and III: Segmental phonetic form. Phonetic intonation

Line II contains a narrow transcription of the sound sequence and the syllable structure of the underlying speech event. Line III represents its intonation (only the pitch contour is given in Table 1), which is the intonation of a speech event that realizes an (elliptic) sentence. By using only one opening and one closing bracket in Line II of Table 1 it is made clear that we are dealing not with three individual phonetic words but with the phonetic form of an entire (elliptic) sentence, in keeping with the fact that in Line III a sentence intonation is represented, not three individual word intonations. In a customary transcription of a phonetic word, it is not only the sound symbols and the syllable boundary markings that appear between the brackets but also stress symbols or names of tones, which both indicate intonational properties of the phonetic word. In Table 1 the intonational properties of the phonetic sentence are represented in a *separate* line because there is no fixed intonation for the phonetic sound sequence and its syllable structure as given in Line II. We readily admit that the marking of syllable boundaries in Line II represents hypotheses concerning phonological words and the properties of their phonetic variants in a sentence context. However, as pointed out in Section 3.3, the hypothetical nature of syllable boundary marking is part and parcel of the hypothetical nature of all parts of the glossing.

4.3 Lines IV and V: Phonological words. Phonological intonation

Line IV identifies the phonological words in a phonemic transcription that is theory-dependent and need not here be explained in detail. It is important to notice, though, that the transcription is of three individual phonological words and simultaneously specifies the sound sequence, the syllables, and the intonational properties (in this case, properties identifying primary and secondary word stress) of each word.

In any syntactic glossing table the actual entries in Line IV depend on the presupposed phonological theory. The existence and character of Line IV does not: Line IV simply

contains the most abstract representations allowed by the phonological theory used. (In a one-level phonology, the representations might be phonetic.)

In some cases we may have non-syllabic phonological words. This is indicated by non-appearance either of a dot (for syllable boundary) or of a syllabicity sign in Line IV, and by "—" in Line V.

Line V represents the intonation at the phonological level, which is a more abstract version of the phonetic intonation given in Line III (once again, only the pitch contour is represented in Line V of Table 1). This more abstract version retains only those phonetic features which are syntactically relevant and, in particular, manifest accent occurrences (there is an accent manifested only on the syllable /blee/ of /pro.ˈblee.mə/, identified in Line V by boldface for the relevant pitch name) or manifest so-called sentence modes (we are dealing with an elliptic declarative sentence, in agreement with the concluding low pitch in Line V and as indicated through use of a period as a punctuation sign in Lines XII and XIII). Regardless of the framework adopted, all modern approaches to accent and sentence mode require a representation of sentence intonation, in particular, of pitch contours, at this level.

As appears from a comparison of Lines IV and V, the pitch contour of the complete phonological intonation cannot be obtained by simply concatenating the word intonations inherent in the phonological words of Line IV. For example, the last syllable of the second phonological word is marked, by the absence of a word stress sign, as having low pitch whereas high pitch appears at the corresponding place of the complete intonation in Line V.

No phonological version is given of the syllabified phonetic sound sequence in Line II, i.e. syllable structure at the phonological level is specified only word-internally. This may create problems in representing sentence sandhi. However, relevant phenomena may also be specified in a Comment part corresponding to Line IV.

4.4 Line VI: Orthographic representation of phonological words

Line VI contains the orthographic names of the phonological words in Line V, using either an established orthography for the language or variety in question or an orthography devised by the researcher on an ad hoc basis. For obvious reasons, it would be unwise to neglect an established orthography in filling in Line VI. On the other hand, no established orthography may be expected to systematically isolate phonological words by means of a one-on-one relationship between orthographic and phonological words. For example, the orthographic name of a form of a clitic may be united with the name of a preceding or following phonological word to form a single orthographic word. This is relevant information on the two phonological words in question and would be given in Line XII (orthographic representation of the entire sentence), if this line is filled in, or in Line XII of morphological glossing tables for the phonological words in question.

Representing relevant orthographic information in the context of Line XII is more informative than simply using hyphens or similar devices in Line VI, apart from the fact that this would complicate the interpretation of the line. Line VI contains orthographic names only of phonological words.

4.5 Lines VII and VIII: Word categories and word form categories

Two types of categories are distinguished: categories that concern complete lexical words (Line VII) and categories that concern only word forms (Line VIII). In one way or another, this distinction is made throughout linguistics, even though it may be terminologically obscured. The categories in question are syntactic not morphological; we are dealing with lexical words and their forms, not with word stems and forms of word stems. The syntactic nature of these categories may be obscured by the category labels used; for example, it may be unclear whether an expression like "Nom" is to refer to all word forms in the nominative or to so-called endings that help identify the nominative word forms ("Nom" for "nominative ending"). It would seem that TGl, too, is unclear in this respect; while only a morphological interpretation is explicitly envisaged, some sort of relation to syntactic categories is apparently also assumed. We do believe that Requirement 2 must be strictly adhered to, which has led us to separate morphological from syntactic glossing in order to avoid all confusion on this fundamental point.

The category listing in Line VII (of word categories) may not yet be complete. It is in Line VII that categories of valency or government would be accounted for. The category labels in both Lines VII and VIII are preliminary and may certainly be changed in a standardization context.

In both Lines VII and VIII (word form categories), only one set of categories is given in each column although several sets of categories would have been possible. For example, "Nom Pl Unm_G Str" was entered in Line VIII for *die*, where "Acc Pl Unm_G Str" would also have been possible. We may require that only category combinations should be represented in a glossing that are relevant for grammatical relations in the given sentence. Even then, there may be more than one category combination that satisfies this condition. Indeed, both "Nom Pl Unm_G Str" and "Acc Pl Unm_G Str" are relevant in the case of our elliptic sentence. To avoid unnecessary duplication, only one category combination should appear in a glossing table, and the others may be accounted for through a part of the Comment correlated with the relevant cell of the table.

4.6 Line IX: Meanings and semantic effects

What is represented in Line IX is (i) two concepts that are lexical meanings of, respectively, the *unübersichtlichen*-part and the *die probleme*-part of the sentence, and (ii) a set of categories that characterize the *die probleme*-part.

a. Lexical meanings

The first of the two concepts is listed in column 2, the *unübersichtlichen*-column, the second is listed in column 3, the *probleme*-column although it is to be associated not just with the *probleme*-part but with the *die probleme*-part as a whole. This association is due to the fact that the *die probleme*-part is treated as a complex form of the lexical word *Problem* (which is a fact expressed in the following Line X), and by any customary conception all forms of a lexical word have the same lexical meaning. Therefore, if in Line IX the concept name is entered in column 3 (where the 'main part' of the complex form is accounted for), this is to characterize the entire *die probleme*-part.

A name of a concept can be suggestive of the content of the concept only to a certain degree. As a rule, there will be differences, be it subtle ones, between the meaning of the concept name in the language from which it is taken (English, in this case) and the content of the concept named. Indeed, the meaning associated with *unübersichtlichen* is a good example: There simply is no English word with precisely this meaning; so choosing "involved" in single quotes as a name of the meaning only yields an approximation. The difference should be spelled out either in a part of the Comment correlated with the cell in question or, since there is a morphological glossing table for *unübersichtlichen*, in a part of the Comment of this table correlated with its Line XIII (see below).

There is also the problem of choosing a specific language for the concept name. We suggest that only one language — normally, English — should be used for concept names in the glossing table, which must then be represented also in Line XIII as the language of a sentence meaning paraphrase. Concept names from other languages may be introduced in the Comment of a syntactic table, of a morphological table, or of both.

b. Semantic effects: the auxiliary part of complex word forms

The entry in Line IX in the *die*-column, "Nom Pl Def", is not a meaning of anything. The entry indicates that the *die probleme*-part, treated as an occurrence of a complex noun form, is syntactically categorized by the three categories Nom (the set of nominative forms), Pl (the set of plural forms), and Def (the set of definite forms, i.e. exactly the forms with a definite article occurrence). Def can be associated with the *die*-occurrence as its semantic effect (in some reasonable sense: occurrence of the syntactic category Def affects the construction of sentence meanings). The remaining categories, Nom and Pl, are in some way determined by the separate categorizations in Line VIII of the *die*-part and the *probleme*-part. Nom and Pl are here listed again in Line IX because the general conditions for obtaining them from characterizations in preceding lines are not immediately clear.

It may be argued that there are complex verb forms in German but no complex noun forms. We do believe that a good case can be made for complex noun forms, too. They are here chosen for their greater simplicity but the treatment of complex word forms in glossings could have been demonstrated just as well, if more laboriously, by means of verb form examples.

In the case of a complex verb form it is customary to distinguish between its *auxiliary part* and its *main part*, a distinction that carries over to arbitrary complex word forms. In our example, the auxiliary part, the *die*-part of the sentence, is extremely simple. As demonstrated by complex verb forms in German or English, this need not be the case. Given an auxiliary part that consists of several phonological words, we obtain several columns in a glossing table, and it may be possible to associate different categories in Line IX with different phonological words in the auxiliary part. In this case, relevant categories should appear in Line IX separately in relevant columns.

c. Many-word word forms without an auxiliary part

A distinction should be drawn between complex forms (forms with an auxiliary part — more precisely, a non-empty auxiliary part) and forms that simply consist of several phonological words. Forms with several phonological words may still not have an auxiliary part. In particular, a form of a circumposition such as German *um–willen* has several phonological words but no auxiliary part. This difference could be brought out by treating

many-word forms without an auxiliary part somewhat differently in a glossing table from complex word forms: in glossing a many-word word form without an auxiliary part, only the column for the first phonological word contains entries in Lines VII to IX. Normally, a concept name will appear in Line IX. It is only in Line X, the constituent structure line, that the various phonological words are characterized as belonging together (see Sec. 5.2, below).

d. Semantic effects: particles

There is one type of semantic effect not yet represented in our example: the semantic effect of particles to which a lexical meaning is normally denied, particles such as the negation particle *nicht* (understood as a lexical word) or the so-called modal particles (*Abtönungspartikel*) that play such a large role in German. There is a lot of disagreement on the semantic treatment of these particles. One possibility would be to associate with them semantic functions, functions that are used in the construction of sentence meanings. A suggestive name of such a function (or whatever the semantic effect of the particle is taken to be) should also occur in Line IX in the appropriate column when we have an occurrence of one of these particles.

4.7 Line XII: Rendering of the sentence in an established orthography

An established orthography for the language or language variety will not always be available, and application of AGl does not depend on it. On the other hand, there should be a line for an orthographic rendering of the sentence, for various reasons. For example, an important ulterior purpose of the documentation may be creation of texts to be used directly by the speech community. Also, the orthographic naming of phonological words in Line VI may have to deviate from the conventions for orthographic words in an established orthography. In our case the third orthographic word in Line VI would have to be capitalized; and in other examples there may also be discrepancies between the number of orthographic names of phonological words in Line VI and the number of orthographic words in the representation of the sentence in an established orthography. All such discrepancies are brought out by a comparison of Lines VI and XII, in case Line XII is filled in (morphological tables may also be helpful, see Sec. 6.7, below). Line XII is not subject to the division into columns because it simply follows the orthography.

While AGl applies independently of any pre-existing orthography, the only texts available may be written ones. In this case, the written texts provide the raw data for glossing, which, as a rule, will mean a direct jump to Line VI in applying AGl. In making this jump, we still assume that we have *written* raw data for an *oral* variety. In the documentation of a *written* variety the entries in Line VI would have to be re-interpreted as names of grapheматic words. The interpretation of Line XII changes accordingly. (Details would have to be worked out.)

4.8 Line XIII: Sentence meaning paraphrases

The only entry in Line XIII consists of an abbreviated language name (“E” for “English”) followed by a colon followed by an (elliptic) English sentence that paraphrases the meaning of the German sentence, which is orthographically named in Line XII. Formally,

Line XIII of a syntactic glossing table is a list where each entry has such a form. This allows for sentence meaning paraphrases in different languages.

Each paraphrase renders at least part of the sentence meaning. Certain parts of a sentence meaning are not easily paraphrased, such as parts that relate to speaker attitude. A sentence meaning paraphrase will therefore, as a rule, require supplementation by entries in the Comment part that is correlated with Line XIII.

Generally, a meaning paraphrase together with its supplements should characterize the sentence meaning as precisely as feasible. Any sloppiness will create problems as soon as the glossings are used to formulate hypotheses on sentence meaning composition in the language or variety being documented.

5 The syntactic glossing table (2): Representing structural and relational information (Lines X and XI)

5.1 General remarks

Existing glossing formats do not yet systematically represent either information on syntactic structure or functional information, in particular, information on the usual grammatical relations. Writing function names like "Subj" into a TGl glossing is sometimes done on an ad hoc basis or for characterizing certain morphemes only. On the other hand, it is obvious from Requirement (1) (glossings as a basis for grammar writing) that structural and functional information must be retrievable from a glossing. Some relevant information is indeed contained in Lines I to IX. However, this is not yet sufficient to identify either the syntactic structure of the sentence or the relations that occur in it. In particular, a constituent structure of some sort should be retrievable from a glossing table. Normally, constituent structures are given by tree diagrams, or by equivalent formulations. Trees clash with the linear nature of glossings. Existing linearizations through bracketing are unwieldy and quickly become uninterpretable once we are dealing with real-life sentences, which may be complex and long. Lehmann, for one, despairs of finding solutions to the representation problem for syntactic structures (1982: §4.7).

We have been experimenting with a strictly linear format for the information that is still missing from Lines I to IX, but is needed for a surface constituent structure of the sentence. Such formats appear to be possible, but involve so much coding and decoding that they are ultimately not worth the effort. We therefore suggest a line in form of a list, and another list for representing relational information, emphasizing the following points:

(i) The lists introduce redundancy into the glossing table by making explicit information some of which is implicit in preceding lines, but then, redundancy-free glossing can hardly be imposed as a general requirement if the documentation format is to be of practical value.

(ii) The information on the syntactic structure and the grammatical relations must in principle be relevant to grammar writing independently of the format used. In a given grammar, the structural and the relational information will then be employed but the form in which it is used will be theory dependent.

It is here assumed — correctly, we believe — that construing Lines X and XI of a glossing table as in the example makes their content useful, even indispensable for the grammatical analysis of the language or variety whatever theoretical framework is adopted.

5.2 Line X: Surface constituent structure

Line X contains three separate entries all constructed on the same pattern: First, there is a digit, or a sequence of digits separated by commas; this is followed by a colon followed by a category name (cf. Sec. 2.2). Each entry can be read as in the following example:

"1,3: Nf" for "the part of the word sequence consisting of the phonological word in column 1 and the phonological word in column 3 is associated with the category Noun form".

The entries form an unordered list and could as well appear in a different order.

In the third entry of Line X, two digits are marked by bold face (some other marking could also be used). This means, intuitively, that the 1-3-part of the word sequence is the nucleus or head of the 1-2-3-part, that is, *die probleme* is the nucleus of *die unübersichtlichen probleme*. The information coded by bold face for digits is no longer information on the constituent structure but is relational information which could have been represented in the following Line XI but is, for practical reasons, embodied already in Line X.

The details of this example are theory dependent with respect to the categories used and in presupposing a specific solution to the DP/NP-problem. The line format is, however, theory independent. The example also demonstrates how complex word forms may be identified: neither "1" nor "3" appear separately each with a category symbol, and there are entries in Lines VII to IX both in column 1 and column 3. (In the case of a many-word word form without an auxiliary part, there would also be digits treated in this way but there would be entries in Lines VII to IX only in the first column relevant to the word form.)

It is an advantage of this line format that discontinuous constituents are represented directly (*die probleme* is discontinuous because of "1,3", where "2" is missing).

From the entries in Line X a tree diagram is easily constructed; and it should not be difficult to write an algorithm for its automatic generation.

5.3 Line XI: Relational information

Due to the simplicity of the example the list of entries has only one item. The item is, however, sufficient to exemplify the format for arbitrary entries: Each entry in Line XI consists of a name of a syntactic relation (usually, a traditional grammatical relation) followed by a dot followed by a sequence of digits or digit sequences; members of the sequence are separated by blanks. Such entries may be read on the pattern of the single item in Line XI:

"mod: 2 1,3" for "the part of the word sequence consisting of the phonological word in column 2 is a modifier of the part of the word sequence consisting of the phonological words in columns 1 and 3".

Informally, *unübersichtlichen* modifies *die probleme.*

Information on the nucleus or head relation in the sentence was represented already in Line X; from a systematic point of view, it should appear in Line XI. If another sentence had been chosen, we might have had entries on other relations, too. In most if not all cases, such entries could be constructed on the basis of information contained in Lines VII to X provided Line VII also contains entries on valency categories.

Even relational ambiguity is not beyond the format adopted for Line XI: A single sequence of digit sequences may combine with different relation names.

Once again, the relations assumed for Line XI in a given syntactic glossing table are theory dependent, but this does not hold of the line format itself; whatever the relations assumed, as long as these are 'surface relations' in a traditional sense whose occurrences are coded by means of numerals as indicated, the line format remains unaffected.

6 The morphological glossing table

6.1 General remarks

Table 2 is arranged in a form strictly analogous to Table 1, and many explanations for Table 1 simply carry over to Table 2. Lines VI to XIII of a morphological glossing table are strictly analogous to the corresponding lines of a syntactic one; the syntactic entities in the syntactic table are simply replaced by corresponding morphological ones.

There is, however, a major difference in Lines II and III: In Table 1, the entries in Lines II and III are phonetic; in Table 2, they are phonological. This is due to the fact that our morphological glossing is for a phonological not a phonetic word, is for a word listed in Line VI of Table 1. The phonological words that figure in a syntactic glossing table are subjected to morphological not to phonetic glossing in a morphological glossing table.

It may seem that this prevents us from providing phonetic information on phonological words when only individual phonological words, outside a sentence context, are available, for example, words contained in a word list. But suppose that we are dealing with raw data that appear to be realizations not of sentences but of individual words. Even in this case, what we are really confronted with is realizations of elliptic sentences. For example, a single word may be realized in answering a question such as, "What is this word?" The answer elliptically realizes the sentence (in English): "This word is . . . ". We may well be interested only in the phonetic and morphological properties of the phonological word that is realized. We would then fill in only the first four lines of a one-column syntactic glossing table, and all lines of a corresponding morphological one.

This characterises the documentation situation from a systematic point of view, touching on various rather subtle and controversial points in phonology. Obviously, in such a documentation situation, shortcuts may and will be used.

The second phonological word in Table 1, *unübersichtlichen*, is chosen as an example for the morphological Table 2. Individual correspondences between the two tables will as a rule not be pointed out.

6.2 Lines I to V: Number and order of morphs — Segmental phonological form — Phonological intonation — Morphs — Morphological intonation

Line I is on the number and order of morphs in *unübersichtlichen*, specified by Arabic numerals; there are five morphs.

Line II gives the syllabified sound sequence of *unübersichtlichen*, and Line III its (phonological) intonation. Lines II and III are jointly equivalent to the entry in cell IV.2 of Table 1 which names the phonological word in phonological notation; Lines II and III differ from this entry only by making the intonational properties of *unübersichtlichen* explicit. It may come as a surprise that in German, too, pitch contours are assumed to be a major part of word intonations. If different assumptions are made, the content of Line III changes accordingly. Making the assumptions on German happens to have a fortunate by-effect for our presentation: It also demonstrates how, in a phonological word of a tone language, tones would be represented explicitly by naming pitches (level pitches or glides).

Lines II and III of a morphological glossing table contain only information that would already be represented in a corresponding syntactic glossing table, even though Line III makes explicit the intonational properties of the phonological word. This may allow for shortcuts in filling in the table.

Line IV differs from Lines II and III in naming not the phonological word but the individual morphs, as indicated by means of slashes before and after each entry in Line IV. (Using slashes in Line IV in this way is vital from a systematic point of view. Naturally, such repetitive features of an entry are obvious candidates for automatization.) It should be noted that the entries in Line IV specify morphs completely, including their intonational properties. In a tone language the word stress symbols would be replaced by tone symbols. So-called free tones in a tone language may be construed — in agreement with their treatment in Autosegmental Phonology — as morphs without a sound sequence and syllable structure, and represented by means of a separate column that has a bar symbol in various lines.

Line III, which gives the pitch contour of the intonation of the phonological word, is not yet sufficient to specify the pitch properties of all morphs named in Line IV. For example, the representation of the *sicht*-part of *unübersichtlichen* in Line II is marked in Line III by "H,L", indicating secondary word stress, whereas the morph *sicht* in Line IV has the symbol for primary word stress. It is only in the context of the entire phonological word that the pitch for primary word stress (H) is replaced by the pitch characterization of secondary word stress (H,L). This shows that the pitch contour of the intonation of *un über sicht lich en* —a 'morphological word'— is obtained from the pitch contours of the individual morphs but need not be identical to the sequence of these contours.

This may appear as a very subtle point specific to one analysis of German word intonation. However, anybody who has ever studied a tone language will immediately remember a basic phenomenon in such languages, namely, the expression of morphological relations by means of systematic changes of the tones of the relevant morphs. Any format for morphological glossing must provide for such phenomena.

What is represented in Line V is the pitch contour for the morphological word *un über sicht lich en*, denoted in an abbreviated way through which Line V becomes identical to

Line III, which names the pitch contour for the phonological word *unübersichtlichen*. For this reason, Line V may simply be left empty, with blanks in the cells.

6.3 Lines VI to VIII: Orthographic representation of morphs — Stem and morpheme categories — Stem form and morpheme form categories

Line VI contains a sequence of orthographic names for the morphs in Line IV. The sequence must agree with the orthographic name chosen for the phonological word in the syntactic glossing table.

In a tone language we may have a non-segmental morph ('free tone'), which may be hard to represent orthographically. If no orthographic name is chosen, a bar sign appears in Line VI in the column for this morph.

Line VII supplies new information of a properly morphological kind: information on the stem and morpheme classes associated with the various morphs. These classes are indicated schematically (see Sec. 2.4). The expression "$SubSt_l/AdjSt_m$" may be read as "the set of morphemes that combine with a form of a substantive Stem of class l to yield a form of an adjective Stem of class m". Line VII presupposes a wide-spread conception in morphology by which a Stem, usually, a so-called word stem, may have several stem *forms*, and is different from any of its forms if only trivially so. The distinction may also be extended to grammatical morphemes.

Distinguishing Stems (word stems) and grammatical morphemes from their forms, we have, in particular, morphological categories for Stems (Line VII) and morphological categories for stem forms (Line VIII). Categories for stem forms are widely assumed, for example, in speaking of Preterite stems (i.e., Preterite stem forms) as opposed to Present tense stems. In Table 2, there are no relevant stem form or morpheme form categories, and this is indicated by a bar in each cell of Line VIII.

The Stem and morpheme categories in Line VII may have to be characterized by also referring to stem form categories; for example, the *bar*-suffix in German combines mainly with forms of Stems of transitive verbs, and these forms must be Present tense forms. Coding such requirements in the name of a Stem or morpheme class is non-trivial and not easily subjected to standardization.

Once again, the details for filling in Lines VII and VIII may be theory dependent; the fundamental distinction between, say, Stem classes and stem form classes is made more or less universally.

6.4 Line IX: Meanings and semantic effects

This Line again specifies lexical meanings and semantic effects. The lexical meanings associated with stem morphs are of exactly the same type as the meanings associated with phonological words.

Among the semantic effects associated with individual morphs, we again have two types. One is exemplified by the effect associated with *lich* in column 4 and denoted by "suitable-for" in Line IX. Intuitively, suitable-for takes a meaning associated with *über sicht*, say, the concept 'overview', and transforms it into a corresponding 'suitability meaning', say, the concept 'suitable for overview'. The precise nature of such a semantic effect depends

on the presupposed theoretical framework. Ontologically, the effect associated with *un* in column 1 and denoted by "not" in Line IX is of the same kind. Intuitively, not takes a meaning associated with *über sicht lich* and 'negates' it. The two effects are typical of derivation affix effects.

The semantic effect associated with *en* in column 5 is given in Line IX as: Unm_C Pl Unm_G Wk. This expression is to be understood exactly as in Table 1, that is, the effect associated with the inflection morph *en* is a set of syntactic, not morphological categories. This is typical of the semantic effects associated with inflection morphs.

The example chosen for Table 2 does not tell us how to deal with morphological intransparency. Suppose that we wish to analyze *sicht* into two morphs, *sicht* and *t*, thus creating one column for *sich* and one column for *t*. This would resume the original derivation of Stems of verbal nouns from Stems of verbs by means of a suffix *t*, which is no longer productive but has left behind a number of semi-transparent stem forms. How are we to indicate that the meaning of *sicht* is 'view' if we associate the verbal concept 'seeing' with *sich* but do not associate the name of a semantic function (or similar entity) with *t*? Moreover, in some cases we may not even be able to find a concept for a stem morph such as *sich*.

We propose the following solution. If a suitable concept for a stem morph such as *sich* is available, a name for the concept is associated in Line IX of a morphological glossing table with a stem morph such as *sich*, in our case, this would be "seeing" in single quotes. (If there is no suitable concept a bar symbol would be entered in Line IX.) The affix morph (*t*) would normally be associated in Line IX with the name for its semantic effect (which can never be a lexical meaning). In place of this name, we now fill in a name ('view') of the concept that is the meaning of the stem and affix morphs together (*sicht*). Since we now have a concept name in a column for an affix morph and not a name for a semantic affix effect, this concept name is now interpreted as naming the meaning of a complex part of the morphological word. The information on the category of the affix morph in Line VII and the structural information in Line X will tell us what this part is.

Cases of derivation by conversion are covered by means of different morphological glossing tables that are linked by means of their Comments and may be compared for purpose of grammar writing.

6.5 Line X: Representing structural information

a. Overall character of Line X

Virtually all morphological frameworks (at least since Nida 1945) explicitly or implicitly provide for constituent structures in the morphological analysis of phonological words. The details vary from one framework to another. We submit that for the morphological analysis of phonological words structural information must be available that is of the same type as the information supplied in Line X of a syntactic table. Therefore, the entries in Line X of Table 2 again form an unordered list, and are of the same kind as the entries in Line X of Table 1, and are read in the same way. For example,

"**2,3**,4: Stf" for "the part of the morph sequence consisting of the morphs in columns 2, 3, and 4 [i.e. *über sicht lich*] is associated with the category Stem form".

Once again, bold face is interpreted separately to indicate the (morphological) nucleus- or head-relation m-nuc: "the part of the morph sequence consisting of the morphs in col-

umns 2 and 3 [i.e. *über sicht*] is the morphological nucleus of the part of the morph sequence consisting of the morphs in columns 2, 3, and 4 [i.e., of *über sicht lich*]".

There are many more entries in Line X of Table 2 than in Line X of Table 1, which correctly renders the fact that the morphological complexity of *unübersichtlichen* is much greater than the syntactic complexity of *die unübersichtlichen probleme*.

The details of the entries, such as the assumption of Affix form or Stem Group as specific morphological constituent categories, are theory dependent; information on (surface) constituent categories in this line is not.

Again, automatic generation of a tree diagram from the entries in Line X should be fairly easy and can use the same algorithm as in the case of Line X in Table 1.

b. Dealing with morphological discontinuity

There is no discontinuity in the case of *unübersichtlichen*, and this is typical of German morphology. However, morphological discontinuity is a basic linguistic phenomenon and must be accounted for. The following proposals are, to the best of our knowledge, in agreement with what is implicit in the descriptive formats used for languages that typically exhibit discontinuity at the morphological level.

There are two types of discontinuity, created, on the one hand, by so-called circumfixes and, on the other, by infixes. A form of a circumfix may be construed as a sequence of two or more morphs (somewhat stretching the usual sense of "morph"), and the occurrences of a circumfix form are dealt with in a morphological glossing table in exactly the way occurrences of a form of a circumposition are treated in a syntactic table, see above, Sec. 5.2. In many cases where infixes could be postulated an alternative treatment via stem form alternation may be preferable but it seems unwise to exclude infixes quite generally. A true infix creates 'split stem forms'. Discontinuity of a stem form can also be treated by construing the stem form as a sequence of morphs, each with its separate column in the morphological glossing table, where only the first column may be filled in completely and where Line X of the glossing table indicates that these morphs belong together.

6.6 Line XI: Relational information

Although not as wide-spread in morphology as the use of constituent structures, morphological relations — largely patterned on grammatical relations in syntax — are included in most contemporary frameworks and should therefore be provided for in morphological glossing.

Line XI in Table 2 has been construed as precisely analogous to Line XI in Table 1, in particular with respect to the form and interpretation of individual entries. The only entries that must be explained are the ones with "m-qual", to be understood as follows:

"m-qual: 4 2,3 3" for "the part of the morph sequence consisting of the morph in column 4 [*lich*] qualifies morphologically the 2,3-part of the morph sequence [*über sicht*] with respect to the 3-part [*sicht*]".

This relation is patterned on syntactic relations such as negation and is here assumed for German for reasons that need not concern us in the present context.

6.7 Lines XII and XIII: Rendering the phonological word in an established orthography — Word meaning

In Line XII the rendering of the phonological word in the established German orthography ("unübersichtlichen") differs from the orthographic name ("*unübersichtlichen*") in Line V of Table 1 only by not being in italics. In other cases, differences may be less trivial. In particular, one phonological may require two orthographical words in the established orthography, and conversely, two phonological words may have to be rendered by one orthographic word. Because of such deviations, a separate Line XII is justified even in a morphological glossing table.

The entry in Line XIII of Table 2, "involved" in single quotes, is identical to the name of the meaning of *unübersichtlichen* in Table 1. Explanations for the name of the meaning (such as "more precisely: 'hard to analyze in all respects' ") may have been given already in the Comment of the syntactic Table 1 but are more naturally introduced in the Comment of the morphological Table 2, in a part of the Comment correlated with Line XIII; this line of Table 2 can be linked to the relevant cell in Table 1.

Appendix 1. Table 1: Syntactic Glossing Table

I	1	2	3
II	[di.	ʔʊŋ̩y.bɐ.zɪç.tˡɪç̊n̩	pʁo.blee.mə]
III	L	H_f M_f L H H H	H H_r L_f
IV	/ ˌdii. /	/ ˈʔun.ˌʔyy.bə.r-ˌziX.t-ˌliX̟ə.n /	/ pro.ˈblee.mə /
V	L	H L L H H H	H **H_r** L
VI	**die**	*unübersichtlichen*	*probleme*
VII	DefArt	Adj	Sub Neut
VIII	Nom Pl Unm_G Str	Unm_C Pl Unm_G Wk	Nom Pl Unm_D
IX	Nom Pl Def	‘involved’	‘problem’
X	1,3: Nf 2: Nf **1**,2,**3**: NGr		
XI	mod: 2 1,3		
XII	Die unübersichtlichen Probleme.		
XIII	E: The involved problems.		

Appendix 2. Table 2: Morphologic Glossing Table

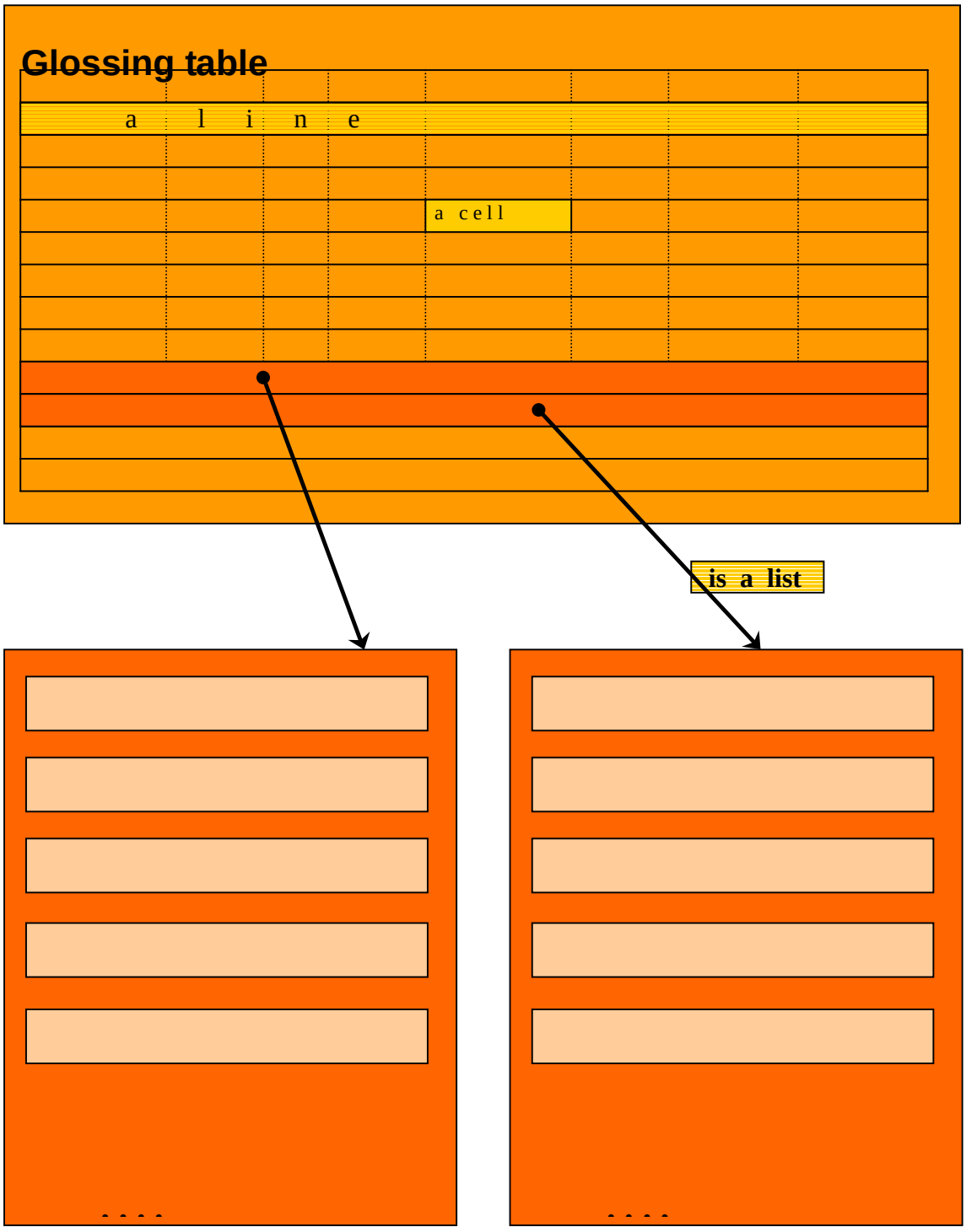
I	1	2	3	4	5
II	/ ʔun.	ʔyy.bə.r-	ziX.t-	liX̣	ə.n /
III	H	H,L L	H,L	H,L	L
IV	/ ˈʔun. /	/ ˈʔyy.bə.r /	/ ˈziX.t /	/ ˌliX. /	/ ə.n /
V	H	H,L L	H,L	H,L	L
VI	*un*	*über*	*sicht*	*lich*	*en*
VII	$Pref_i$	$PrepSt_j$	$SubSt_k$	$SubSt_l/AdjSt_m$	$AdjFlex_n$
VIII	–	–	–	–	–
IX	not	‘over’	‘view’	suitable-for	Unm_C Pl Unm_G Wk
X	1: Af 2: Stf 3: Stf 4: Af 5: Af 2,**3**: Stf **2,3**,4: Stf 1,**2,3,4**: Stf **1,2,3,4**,5: StGr				
XI	m-mod: 2 3 m-qual: 4 2,3 3 m-mod: 1 2,3,4 m-qual: 5 1,2,3,4 1,2,3,4				
XII	unübersichtlichen				
XIII	‘involved’				

Appendix 3.

Overall structure of the format

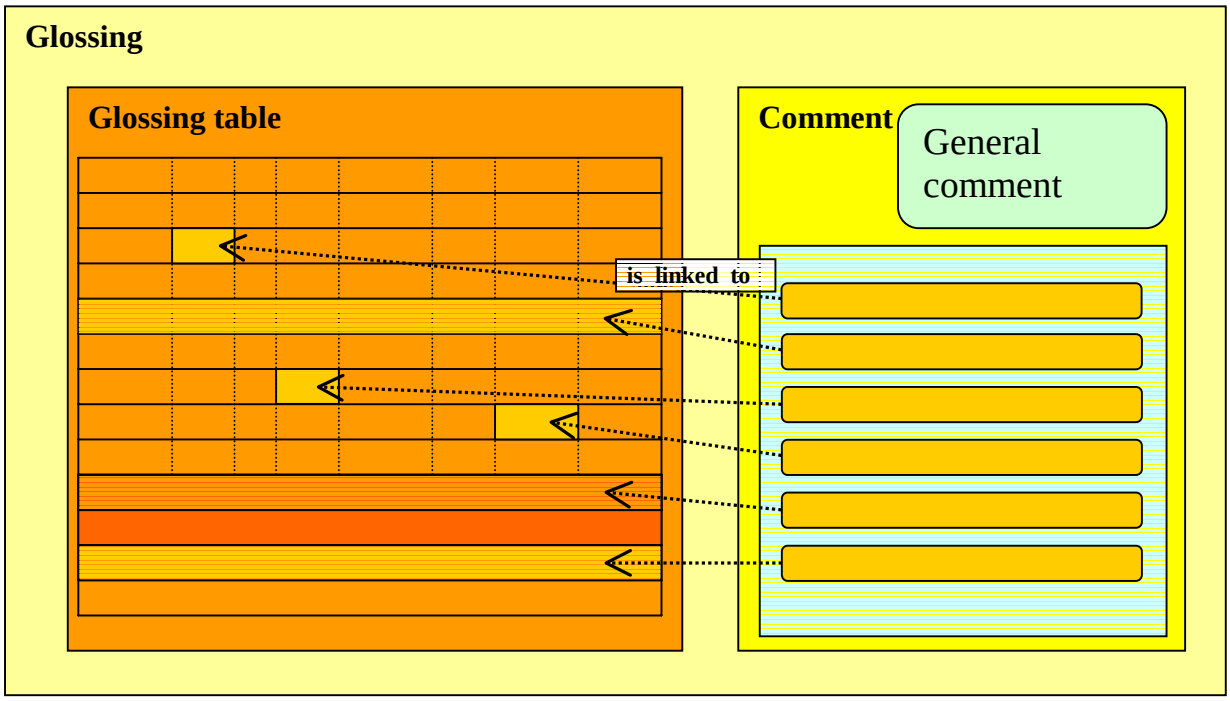
Dotted vertical lines are added to make the cell structure of the tables more obvious. Dots (“. . . .”) indicate an unspecified number of entities such as items of a list.

1: A glossing table

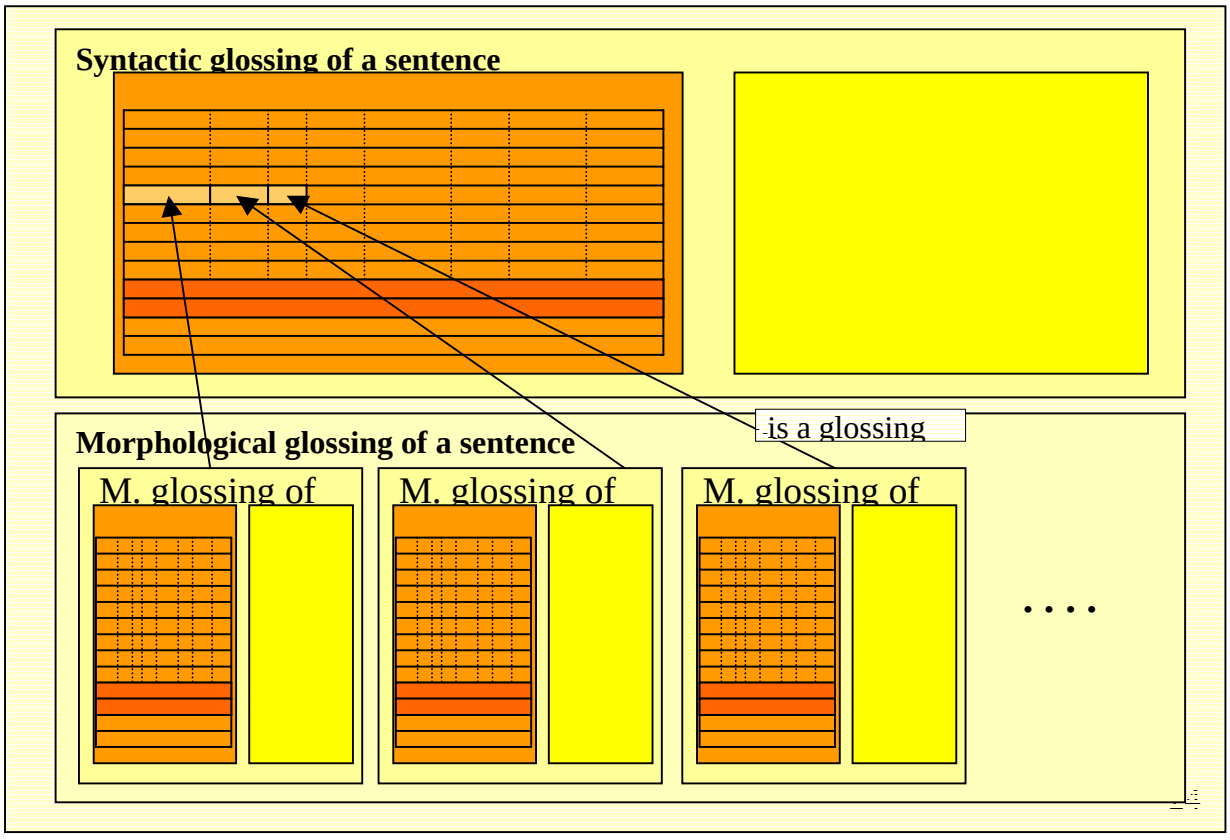


2: A glossing

(i.e., a syntactic glossing of a sentence or a morphological glossing of a word)



3: Morphological and syntactic glossings of a sentence



4: A documentation of a text