Realizing Humboldt’s dream: Cross-linguistic grammatography as data-base creation

Dietmar Zaefferer

1. The idea of the Cross-linguistic Reference Grammar database

1.1. Humboldt’s dream

In his *Einleitung in das gesamte Sprachstudium* (‘Introduction to the general study of language,’ 1810/11), Wilhelm von Humboldt complains about the lack of a general comparative grammar and the abundance of judgments that all too obviously lack the firm ground of properly established leading ideas. Almost two hundred years later we know much more about language, but the project of a general comparative grammar based on the firm ground of properly established leading ideas is still far from completion and the issue has become much more urgent given the increasing speed of extinction of languages in the course of globalization. Fortunately, there are several projects under way that are working on this huge task. One of them is CRG, the Cross-linguistic Reference Grammar database project. One way of describing its basic idea is that it strives to make Humboldt’s dream come true. The CRG enterprise goes back to a joint initiative of Bernard Comrie, Bill Croft, Christian Lehmann and the author of this contribution at the end of the eighties and in the beginning nineties (cf. Comrie et al. 1993, Zaefferer 1998). The aim was to create some kind of revised electronic version of the famous Lingua descriptive studies questionnaire (Comrie and Smith 1977), a framework for the description of human languages of any type.

Any realization of Humboldt’s dream has to come to grips with three fundamental problems:
The comparability problem

The typological bias problem

The theoretical bias problem

In order to get an idea of the first problem (the other two will be taken up in the subsequent sections), imagine a library with about 7000 grammars of the world’s languages including the Routledge Descriptive Grammars (DG), the publications of the Mouton Grammar Library (MGL) and many, many others. That would be a rather impressive accumulation of linguistic knowledge and an ideal place for cross-linguistic research. But when one thinks about doing practical work, it quickly turns out that even in that ideal place it would be rather time consuming to reach reliable conclusions about the diversity or uniformity of so far existing manifestations of the human language faculty.

First one might object that for many purposes such as the discovery of preferences or statistical universals, there is simply too much information: In such a context an exhaustive search is not necessary and a representative sample would be sufficient. But this is not a real objection because having too much information is not a problem as long as it does not interfere with quick accessibility.

The real problem is still lack of information: Even the incredible amount of information in that imaginary library is not enough to ensure reliability of cross-linguistic work since what is missing is metainformation: As long as it is not clear that the terminology used in grammar $A$, say Evans’ MGL grammar of Kayardild, is the same as the one used in grammar $B$, say Fortescue’s DG description of West Greenlandic, i.e. that they are based on the same cross-linguistically valid operationalizations, there is plenty of room for misanalysis. Both faux amis (ambiguity: use of the same terminological label for different concepts) and faux ennemis (synonymy: use of different labels for the same concept) occur again and again and are a big obstacle for the proper comparison of languages.

Another kind of metainformation is organizational in nature: How are all the examples and partial descriptions organized into a whole? This problem has been solved in the DG series by the use of a common table of contents, but not in the MGL volumes. The latter have the advantage that they are able to evolve but pay the price of problematic comparability, the former have the advantage of high comparability at the expense of being outdated in their structure (cf. Comrie 1998). This seems to be a dilemma that cannot be solved in the realm of paper grammars and that therefore requires
a migration to electronic grammars (for a comparison of the two cf. Zaef-
ferer 1997), where updates and reorganization are much less expensive.

1.2. Advantages of an electronic database format

The first and probably most important advantage of an electronic database
format over a paper grammar is its flexibility. The Comrie/Smith question-
naire of 1977 could of course be changed, but at the expense of comparabil-
ity: All grammars that have appeared so far would have to be reprinted with
the new table of contents and with added or rearranged information. The
latter problem persists with the electronic format, but improving on the
structure is much easier and no reprinting is necessary although printouts
are of course always possible. As Comrie (1998: 14) rightly points out it is
important that new advances in linguistic theory and our understanding of
cross-language variation can be more readily incorporated into the frame-
work and that existing descriptions can be rather easily supplemented in the
relevant respects.

The second advantage lies in the organization of the information in the
database. Here the CRG project has opted for a systematics or ontology of
linguistic phenomena that is mainly based on two kinds of relations, taxo-
nomic and meronomic ones. Therefore the database structure has the form
of a tree with three kinds of edges: BE-edges for the taxonomic relations (a
negative clause is a clause), HAVE-edges for the meronomic relations (a
negative clause must have a negation marker) and optional edges for the
rest (a negative clause may have a secondary negation marker).

The third advantage lies in the relative independence of the way the in-
formation obtained from the database is displayed from the organization of
the information in the database. Whereas the information in, e.g., the De-
scriptive Grammars is primarily organized by languages, and only secon-
darily (via the shared table of contents) by linguistic subsystems and phe-
nomena, a database can handle this ‘vertical’ kind of organization with the
same ease as a ‘horizontal’ kind of organization where, e.g., vowel invento-
ries are described across languages and then consonant inventories. With
the migration to an electronic database difficult publication decisions
(grammars of entire languages versus comparative studies of subsystems)
turn into simple display options which can be ranked at will.
2. Language description and the challenge of general comparability

Some theoretical linguists tend to look down at mere language description. Explanatory adequacy is more ambitious than descriptive adequacy, therefore it seems to be more interesting. But that does not mean that descriptive adequacy can be neglected. On the contrary: It is pointless to explain data taken from a descriptively inadequate grammar. And until today the number of really good descriptive grammars and therefore the number of really well-described languages is very limited.

But whenever one speaks of standardizing language descriptions across language types, taking up the challenge of general comparability, the spectre of Procrustes enters the room and scares the discussants. Procrustes, as is well known, used brute force in making different customers alike, stretching the short ones and cutting off the feet of the taller ones. Wouldn’t a general framework for language description be like Procrustes? Here the typological bias problem mentioned in section 1.1. above raises its ugly head: Any general framework for language description has been developed on a rather small sample of languages. So chances are this sample is typologically biased and therefore the framework is helpful for the description of languages that conform with the bias and it may be more of a hindrance for the description of other languages. The only solution to the problem is to set up the sample used in the framework development in such a way that it contains languages that are maximally apart from one another in the different dimensions of typological variation. The example presented in section 6 below illustrates such a long distance comparison in the dimension of analyticity/syntheticity. It is left to the reader to judge if the descriptive apparatus is biased towards one language or the other.

Once the typological bias problem is solved, a more adequate picture should replace the one of Procrustes. It would be the picture of a person who simply measures. So instead of saying these two languages are rather similar in a given respect, say their vowel inventory, one could come up with an objective value on a scale.

One problem with the comparative description of languages so far is the following: There are what may be called familiarizers, linguists who try to show that all languages are basically the same and therefore tend to play down the differences, and there are exoticizers, linguists who try to prove that their language is incommensurable with all others and therefore tend to play down the commonalities. Both attitudes are of course exaggerations,
but without measurements it will be hard to find a non-arbitrary balanced position between the two extremes.

Once the challenge of general comparability has been taken up and the prospects for an adequate treatment are good, it makes sense to ask: What is a general comparative grammar? According to the conception adopted for the CRG project, a General Comparative Grammar is a grammar that describes each phenomenon of each individual language by assigning it its systematic place in the typological space, i.e., the universal space of possible linguistic phenomena. Simply by being assigned its place in this space each phenomenon is automatically compared with all other phenomena in it. The space of possible linguistic phenomena is an $n$-dimensional space, where $n$ is the number of dimensions of variation in which linguistic phenomena can be assigned a place (value). It is assumed that lexical phenomena (in the sense of the static lexicon) can be characterized by a set of finitely valued basic dimensions of variation and that characterizations of phenomena defined by the grammar proper (dynamic lexicon, i.e., productive word-formation, and syntax) result from adding to this set two further (infinitely-valued) dimensions of variation: word complexity and phrase complexity.

3. The theoretical background of language description

3.1. Why boring theoretical assumptions are more interesting than interesting ones

The third problem stated in section 1.1. above is the theoretical bias problem. One may ask: Why is theoretical bias a problem at all? Why not pick one’s favorite theoretical framework and start from that? Wouldn’t that have the nice side-effect of promoting this framework and giving it an advantage over its competitors? Reflections on the theoretical underpinnings of CRG have revealed an interesting paradox: Strong and interesting theoretical assumptions are good for advancing our understanding of human languages and the forces that have shaped them and continue to make them what they are. But they are not good as a basis for describing linguistic data and the framework that has been chosen for this purpose has no advantage over its competitors.

On the contrary: No ambitious explanatory theory should strive to be included in the theoretical basis of a cross-linguistic reference grammar data-
base. Why? Because explanatory theories are empirical theories and empirical theories strive for falsifiability. But it is impossible to find data that falsify a theory whose assumptions are built into the very description of that data. To give a simple example: Whoever bases his descriptions on a theory that postulates that every sentence has a subject is unable to describe subjectless sentences.

So a high-ranking maxim for the development of the theoretical background of the CRG project was the following: Whenever there is any hope that a hypothesis can be tested against the data in the database, its validity should not be assumed by the theory underlying the database.

3.2. General assumptions

The comparability of human languages is based on their rough functional equivalence: No signalling system qualifies as a language in the intended sense, if it does not provide its users with the means for making assertions, asking questions, making requests, predicking, restricting, modifying etc. Except for the distinction between grammatical and semantic components of the inferable content nothing should be stipulated by the basic assumptions. Among other assumptions the question of the universality of the verb-noun distinction (cf. e.g. Broschart 1997) should be left open.

In his monograph The Rise and Fall of Languages (1997: 132) Bob Dixon speaks about what he calls Basic Linguistic Theory, “the fundamental theoretical apparatus that underlies all work in describing languages and formulating universals about the nature of human language.” I am not sure if such a theory can be identified or whether this is just a cover term for a wealth of overlapping but different sets of basic assumptions used in the work of ‘theory neutral’ descriptions in linguistics. In this respect I am more skeptical than Matthew Dryer (this volume), but the preceding section should have made clear that I agree with Dryer in emphasizing the importance of distinguishing between descriptive and explanatory theories and that for the former something resembling standard practice can be useful. I think, however, that it is useful to not just appeal to such a practice, but to spell out the assumptions, shared or not, which one takes for granted in the descriptions one is making.
3.3. Special assumptions

The following basic assumptions and terminological stipulations are currently in use in the CRG enterprise; they are taken to be acceptable for a broad majority of linguists:

1. Every human language is a system that defines and thus provides its users with an open set of linguistic signs. The openness of this set implies that only some of them can be memorized, while others have to be constructed and interpreted on the fly.

2. A linguistic sign is an abstract conceptual entity, the pairing of a reproducible perceivable form together with a conventionally associated inferable content.

3. Each token of a linguistic sign is therefore a concrete situated instantiation of such a concept: a perceivable instantiation of the form produced together with an inferable instantiation of the content. (The concept of a written linguistic sign token is a little more complicated and will not be spelled out here.)

4. A linguistic sign is complex if components of its perceivable form correspond to components of its inferable content, else it is a simple sign or a simplex. (Thus German rot ‘red’ is a simplex whereas Röte ‘redness’ has two components one of which, the nominalizer, displays a discontinuous shape (umlaut and suffix). It is a matter of debate whether the noun Rot is complex or simple: if analyzed as containing a zero nominalizer it is complex, but if treated, together with the form-identical adjective, as a polycategorial sign it is simple.)

5. It is a ‘fundamental design feature’ (Talmy 2000: 21) of human languages that they have two interlocking subsystems, the grammatical and the lexical, and it is therefore helpful to distinguish between grammatical (form-related, structural) and semantic components of the inferable content of a linguistic sign. Semantic components are conceptual categories whereas grammatical components are formal categories; the latter are either semantically anchored, i.e. indirectly related with conceptual categories (e.g. number with cardinality), or purely formal (e.g. inflexion classes). (Thus the inferable content of the English adjective red has both semantic and grammatical components – the conceptual category of the corresponding color and the semantically anchored grammatical
category adjective – whereas the inferable content of the English suffix -ness of redness has only a grammatical component: that of the semantically anchored grammatical category deadjectival nominalization.

6. The encoding use of language (henceforth ‘linguistic action’) is the situated production of linguistic sign tokens, i.e. the production of perceivable form tokens together with inferable content tokens. The situation in which a linguistic action is performed will be called the production situation. Only the current speaker can be in a production situation. Aspects of the production situation like identity of speaker and addressee, time and location are substantial factors in the determination of the inferable content token.

7. The decoding use of language (henceforth ‘linguistic understanding’) is the situated interpretation of linguistic sign tokens. It consists in perceiving a situated form token and inferring from this a situated content token. The situation in which this linguistic understanding is performed will be called the interpretation situation. Any speaker of the relevant language can be in an interpretation situation. Aspects of the interpretation situation can be factors in the determination of the inferable content token only insofar as they can be anticipated by the respective speaker.

8. The production of a linguistic sign token is a complete linguistic action or speech act only if its situationally inferable content includes a specification of the conventional role it is intended to play in the interpretation situation (illocutionary force). A complete linguistic action is called an illocution.

9. In addition to coding illocutionary force most illocutions (but for instance not uses of interjections) include a propositional act or at least an act of reference.

10. Every human language $L$ defines an open set of linguistic signs whose inferable content includes a semantic component. These signs can be used as the core of a predication; they will be called predicative signs of $L$. The semantic component of their inferable content will be called a predicative concept of $L$. If an entity instantiates such a concept it is predicatively characterized by any corresponding sign of $L$. (The letters you are reading now are predicatively characterized by the English adjective and predicative sign black because they instantiate the coded color concept, which is a predicative concept of English.)
11. Every human language $L$ defines a relatively small closed set of linguistic signs with an inferable content that lacks a semantic component and consists entirely of semantically anchored grammatical content. These signs cannot be used as the core of a genuine predication, and they will be called sortal signs of $L$. The semantic anchor of the relevant grammatical component of their inferable content will be called a sortal concept of $L$. If an entity instantiates such a concept it is sortally characterized by any corresponding sign of $L$. (Consider the letters in these parentheses. They are sortally characterized by the English pronoun this sentence starts with because of the sortal concept of plural cardinality that anchors the grammatical component of plural number included in its inferable content.)

12. Every human language provides means for predicating, i.e. signs that in a production situation include in their inferable content both a predicative concept of that language and a grammatical component that relates this concept to a given entity called a referent. Linguistic signs used for predicating characterize their referent predicatively (and in general also sortally). An act of predicating will also be called a predication.

13. Every human language provides means for referring, i.e. signs that in a production situation include in their inferable content a referent. A possible referent in a production situation is any entity that is either already accessible for predication in the intended interpretation situation or is made accessible by the speaker by introducing it into that very situation. Therefore, acts of reference are either acts of referent maintenance or acts of referent introduction. Linguistic signs used for reference characterize their referent sortally and often also predicatively. This will be called restricting characterization of referents or restricted reference.

14. Every human language provides means for quantifying, i.e. signs that in a production situation include in their inferable content a generalization over virtual referents. Therefore, quantifying sign tokens are non-referential. Like referring signs they characterize their referents sortally and often also predicatively. This will be called restricting characterization of virtual referents or restricted quantification.

15. The combination of a predication with an act of either referring or quantifying yields an elementary propositional act. The inferable
content of a propositional act is called an *elementary propositional content* and consists of the situation it is about (global reference, frame) a situation type. The framing situation is called the *object situation*, whereas both the production situation and the interpretation situation are called *metasituations*. The object situation is the location of the referent of referential sign tokens and the domain of quantification of non-referential ones. If the object situation is specific enough to make the corresponding propositional content true or false, then it is an *elementary proposition*, else an *elementary near-proposition*. An elementary proposition is called *categorical* if it includes an act of referent maintenance, else it is called *thetic*.¹⁰

16. Elementary propositional contents cannot occur alone. They must be embedded either directly or via embedding higher propositional contents in an *illocution* (cf. 8 above).¹¹ The propositional content of an illocution is always the highest propositional content it includes. Assertive illocutions require propositions as contents and are true if their proposition is true, directive illocutions are pointless in that case,¹² other illocution types like erotetic illocutions (questions) require near-propositions as contents and are therefore neither true nor false.¹⁴

4. **Some corollaries**

4.1. The primacy of onomasiology

In section 3.2. it has been stated that the comparability of human languages is based on their rough functional equivalence and that form-related issues like the universality of the verb-noun distinction should be left open. Assuming that the prototypical use of verbs lies in predication and that of nouns in restricted reference, the lack of such a distinction would amount to having a single category of predicate (as in standard predicate logic). According to assumption 13 such a language would simply lack the possibility of an immediate predicative restriction of its referents, i.e. something like a relative clause or desentential attribute would be needed for that purpose.

Generalizing from considerations like this one readily comes to the conclusion that for cross-linguistic grammatography the semasiological (decoding) and the onomasiological (encoding) perspective are not on a par, but that the latter has priority over the former: If comparison is based on
assumptions like ‘there must be a way of expressing roughly this content’, it is safe, but if it is based on assumptions like ‘there must be a copula or a noun-verb distinction’, it is not. Unfortunately, as Ulrike Mosel (this volume) points out, existing reference grammars are almost exclusively organized from a semasiological perspective.

### 4.2. The inseparability of grammatography and lexicography

The primacy of onomasiology also explains why CRG is conceived as an integrated lexicogrammatical database. A look at some coding means for the concept ‘causation of the state of being dead’ shows that they can be found in very different parts of a systematic language description:

1. **English** *kill* in the simplexicon (monomorphemic signs)
2. **German** *töten* in the d-complexicon (derived polymorphemic signs)
3. **German** *totmachen* in the c-complexicon (compound polymorphemic signs)
4. **English** *let die* in the phrasicon (phrasal signs)

(1)-(3) should be listed in the lexicon, (2) and (3) also in the word formation part of the grammar (dynamic lexicon), and (4) in the syntax part of the grammar. A separation of grammatography and lexicography and hence of grammar and lexicon would impede insight into potentially important connections.

### 5. The interlinear representation format (IRF)

The starting point in developing the example data structures of CRG was the tradition of interlinear morpheme glossing combined with translations that is familiar from typological literature, but it was clear from the outset that not only the descriptive, metalinguistic vocabulary is in urgent need of standardization, but also that a much richer structure is not only desirable but also feasible as soon as the data are represented in an electronic format. A richer structure means among other things more lines and so a structure emerged that partly reminds one of a musical score, where the different lines or parts are synchronized by bars. (Mismatches of syllable and morpheme boundaries resemble then a syncopated rhythm.)
The basic dividing line in such an interlinear representation format corresponds to the very notion of a linguistic sign and separates the signifier or perceivable form (above) from the signified or inferable content (below). Here is the IRF-structure of a spoken language sign:

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>+7</td>
<td>audiovisual data (recording)</td>
</tr>
<tr>
<td>+6</td>
<td>coding of <strong>co-linguistic elements</strong></td>
</tr>
<tr>
<td>+5</td>
<td>phonetic transcription of linguistic elements</td>
</tr>
<tr>
<td>+4</td>
<td>representation of higher-level suprasegmentals (intonation etc.)</td>
</tr>
<tr>
<td>+3</td>
<td>autosegment representation (tones etc.)</td>
</tr>
<tr>
<td>+2</td>
<td>phonological segment and syllable representation</td>
</tr>
<tr>
<td>+1</td>
<td>morphophonemic representation</td>
</tr>
<tr>
<td>-1</td>
<td>morpheme gloss with <strong>GRAMMATICAL</strong>, <strong>semantic</strong> and <strong>co-linguistically induced</strong> components</td>
</tr>
<tr>
<td>-2</td>
<td>higher morphological structure</td>
</tr>
<tr>
<td>-3</td>
<td>syntactic structure</td>
</tr>
<tr>
<td>-4</td>
<td>meaning structure (with co-linguistically induced elements in <strong>boldface</strong>)</td>
</tr>
<tr>
<td>-5</td>
<td>literal translation into quasi-English</td>
</tr>
<tr>
<td>-6</td>
<td>free English translation</td>
</tr>
</tbody>
</table>

(The justification for having three different levels of meaning representation, meaning structure, literal and free translation, will be given in connection with an illustrative example in the following section.)

The IRF-structure of written linguistic signs has the same negative levels but the positive levels have a simpler signifier structure:
Realizing Humboldt’s dream

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>+I</td>
<td>+III (or +II, if non-empty) with morpheme boundaries</td>
</tr>
<tr>
<td>+II</td>
<td>empty, if +III is roman, else transliteration of +III into roman-based orthography</td>
</tr>
<tr>
<td>+III</td>
<td>standardized representation of original script with coding of co-linguistic elements</td>
</tr>
<tr>
<td>+IV</td>
<td>reproduction of writing with included elements (illustrations) and including situation (wall)</td>
</tr>
</tbody>
</table>

The IRF-structure of sign language signs is of course a matter of much debate, but it seems plausible that basically the same structure as for spoken language signs can be used.\(^{15}\) (It has to be admitted, though, that in both cases the notation of paralinguistic components remains a challenge.) Here are the representation levels for a sign language sign:

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>+I</td>
<td>morphophonemic representation</td>
</tr>
<tr>
<td>+2.S</td>
<td>phonological representation of strong hand sign components</td>
</tr>
<tr>
<td>+2.W</td>
<td>phonological representation of weak hand sign components</td>
</tr>
<tr>
<td>+3</td>
<td>phonological representation of mouthings</td>
</tr>
<tr>
<td>+4</td>
<td>representation of non-manual sign components</td>
</tr>
<tr>
<td>+5</td>
<td>phonetic transcription of linguistic elements (HamNoSys)(^{16})</td>
</tr>
<tr>
<td>+6</td>
<td>coding of co-linguistic elements</td>
</tr>
<tr>
<td>+7</td>
<td>audiovisual data (recording)</td>
</tr>
</tbody>
</table>

Since there is no widely accepted counterpart for IPA in sign linguistics, the choice of the phonetic transcription system must be more or less arbitrary. The Hamburg Sign Language Notation System, HamNoSys, has the advantage of being highly iconic and not referring to national diversified fingerspelling. It is, however, not very well developed with respect to non-manual components and does not include an adequate representation of mouthings, visible elements from spoken language articulation accessible for lip-reading, which need to be represented on a separate level. Mouthings are important for an adequate representation of many sign languages including DGS (German Sign Language) or JSL (Japanese Sign Language). An interesting difference between the two modalities shows up in the relation between the level +1 and higher levels: Since signs may include sev-
eral simultaneously articulated morphemes, at level +1 an arbitrary linearization of these morphemes has to take place.

6. An example of cross-linguistic description

For the following illustration of the idea of cross-linguistic description two typologically diverse languages have been chosen, languages that are rather far apart on a very prominent dimension of linguistic variation, that of analyticity. Polysynthetic languages, those with polymorphemic and very complex word forms, rank very low on this scale while other languages where the ratio of morphemes to word forms comes close to one rank very high. The former type is represented here by the Iroquoian language Seneca of East Northern America, the latter type is frequently illustrated by Chinese. German is often said to be between the extremes, but the example will show that it is much closer to the analytic end of the scale than to its opposite.

An obvious way of comparing languages is to look for translation counterparts. In the database this means to select level -6 (free translation) in an example and to look for other examples with the same content of level -6. (Looking for literal translation counterparts or matches on even higher negative levels such as meaning structure is possible as well, but will yield rapidly decreasing results.) The Seneca example is taken from a conversation recorded by Wallace Chafe in 1957 (personal communication). The first part of that conversation is about a fisherman who catches a lot of bass. So the third person anaphor relates to this fisherman and the corresponding reference act is an act of referent maintenance in the sense of assumption 13 above. Comparing its description with that of the German translation reveals the commonalities as well as the differences between the two roughly equivalent linguistic signs.
Seneca

+2  deʔ, shos.doʔ, ʔo.ʔh
+1  teʔ- -s- -ho- -stʔshr- -o- -h- - ʔh

-1  NEG- -REP- -3.S.M.PAT- -hook- -be.in.water- -CAUS- -STAT
-2  [PRFX- [PRFX- [PRFX- V.BASE -] -SUFF]]]
-3  [  ]

-4  a  [ILL.TYPE(a):  ASSERTIVE
PROPOSITION(a): < x,  
NEG
(REPETITIVE
(p [REFERENT(p): x [s,ACCESSIBLE, NON-s, SAP, SIMPLEX, MALE.PERSON]
PREDCATE(p): BE.EXPONENT
(e, [ASPECTUALITY(e,): STATE
TYPE(e,): CAUSING(e,): be.located.in.water
EXONENT(e,): y [hook]])])])>

-5  It is not the case that again he is in the state of causing a hook to be in water.
-6  He doesn’t have his hook back in the water yet.

German

+II  Er hat seinen Haken noch nicht wieder im Wasser
+2  er hat saŋ, naŋ 'ha:kan naŋ ničt 'vi:dar im 'vašar
+1  er  hab -t saŋ- -an ha:kan

-3  [[SBJ] [PRED.ROLE] [ ] [DIR.OBJ] ]
naŋ ničt 'vi:dar in-m vašar
CONT NEG REP in.P-DEF.S.N.DAT water.N.S.N.DAT
PRTCL PRTCL PRTCL [PRED.EXT.LOC ]]

-4  a  [ILL.TYPE(a):  ASSERTIVE
PROPOSITION(a): < x,  
NEG
(REPETITIVE
(p [REFERENT(p): x [s,ACCESSIBLE, NON-s, SAP, SIMPLEX, MALE.PERSON]
PREDCATE(p): BE.AGENT
(e, [TYPE.of(CORE): disposing.of.entity.located.at z
UNDERGOER(e,): y [POSSESSED.BY(x, SIMPLEX, hook)])]])>)

-5  He has his hook still not again in the water.
-6  He doesn’t have his hook back in the water yet.
The comparison shows first a striking difference in compactness: The Seneca version has seven morphs coded in five syllables, whereas the German translation requires twice as much effort: thirteen syllables to code twelve morphs. In terms of word tokens the difference is even more striking: The Seneca sentence consists of one (phonological and grammatical) word form, the German counterpart of nine. Only three of the German word forms are complex. Level -2 (higher morphological structure) has therefore been omitted in the German example.

Another interesting difference is in the division of labor between grams with their sortal concepts (marked by small caps) and lexemes with their lexical concepts (marked by italics): Seneca combines eight sortal concepts with two lexical ones, German has four lexical concepts associated with four different part of speech categories and many grams and sortal concepts that occur more than once due to several cases of agreement. This shows a higher degree of redundancy and abstractness: The grammatical features third person and singular on the German verb form help only identify the referent of the predication, the corresponding sortal concepts have nothing to do with the predicate concept itself. Only two of the six occurrences of gender features in the example reflect a sortal concept: The feature ‘masculine’ both in the personal pronoun and in the stem of the possessive pronoun restricts the respective referent by characterizing it as a male person, the other four occurrences of ‘masculine’ (with Haken ‘hook’) and ‘neuter’ (with Wasser ‘water’) are purely formal.

Both sentences are declarative sentences coding an assertive illocution with a categorical proposition as content. The German syntax shows a prototypical transitive construction with a subject, a direct object and a predicate consisting of a finite verb and a locational extension having the form of a prepositional phrase. Three particles modify the propositional content.

The core of the Seneca word structure is the intransitive locational root ‘be in water’ that is transitivized by the causative suffix and intransitivized again by incorporating the nominal root ‘hook’. The whole stem or verb base is then marked as a stative predicate by a corresponding suffix. Therefore, the referential person prefix appears in the patient form since states do not trigger agent forms in Seneca (they are, as it were, not active enough). Two additional prefixes code propositional aspectuality and polarity.

These two different grammatical structures result in different semantic structures. The illocution type ‘assertive’, unmarked in both languages, is in both cases the same, but the proposition structures are different: In the Seneca sentence, the proposition is first negated and then aspectuality-marked
as repetitive, coding the concept of ‘not again’, in the German counterpart aspectuality comes first (noch ‘still’), negation (nicht ‘not’) second, and another aspectuality marker third (wieder ‘again’), coding the concept of ‘still not again’.

The reference parts of the propositional cores are the same, but the inner structure of the predicates differs. The Seneca predication characterizes its referent as the exponent of some stative causation eventity that is indirectly characterized as having the effect that some hook is located in water. So it decomposes the predication core into two causally related subevents with one participant each.

The German predication is different in characterizing its referent as the agent of a semantically transitive, i.e. relational, eventity whose undergoer is some hook that is functionally related (as possessum) to the referent of the predication. So the relation between the person and the hook is not mediated by their participation in two related eventities, but directly expressed in their participation in the same eventity, albeit in different roles. Here complexity shows up in a different place: The relational eventity of having (roughly ‘being able to control in one way or the other’) that serves as predicate core is specified by the location where the control is exerted, serving as predicate extension.

There are of course many details that can be debated and analyzed in a different way, but the basic idea of fine-grained language comparison that such a framework makes possible should have become clear.

The example presented above can be used at several locations in the grammar, e.g., in the chapter on negation, but also under aspectualities, elementary declarative sentences etc.

Just one final remark on the functionality of the database: The type of query chosen for the illustration, i.e. looking for translation equivalents is of course not the only one that is possible. In principle all lines in the IRF, single ones or combinations thereof, can be selected and matching complete examples can be obtained. So if one is interested in higher morphological structures with the pattern \([\text{PREF} - \text{PREF} - \text{PREF} - \text{VERB}.\text{BASE} - \text{SUFF}]]\) one can search for this as well as for declarative sentences in general (with the syntactic structure \([ \text{DECL}]\)).
7. Outlook

Currently the system is being tested in the Munich International Doctoral Program in Linguistics (LIPP) ‘Language Theory and Applied Linguistics’. A guiding principle in its development is the double responsibility for each cell in the database, that results from two orthogonal dimensions of its organization: The vertical dimension is the traditional one that takes care of the consistency and coherence of the descriptions of different subsystems of one and the same language. Responsibility for a column in the database thus corresponds to the notion of authorship of a traditional grammar.

The horizontal dimension is the cross-linguistic one and it is specific to the CRG enterprise. It is here where the balance has to be held between over- and understating the differences and commonalities. Those who are responsible for the rows ought to be specialized in the corresponding subfields like segmental, suprasegmental and phrasal phonology, static and dynamic lexicology, grammar in the sense of syntax and inflectional morphology, decompositional and compositional semantics and pragmatics.

The division of labor that is reflected in this grid structure of the database makes it easy to predict conflicts between a column-related language internal view and a row-related cross-linguistic perspective on the same phenomenon. It is an important advantage of the electronic format that it supports a policy of not only permitting but even encouraging competing descriptions of the same phenomenon (for arguments why this is an advantage cf. Zaefferer 2004).

A central motivation for the CRG project is the firm conviction that the very notion of comparability is of prime importance for the mutually stimulating effects that developments in language description and advances in theory formation could and should have for one another.

A Popperian sequel to Humboldt’s dream will be the use of the database query system as a means for finding counterexamples to any theory that is compatible with the one proposed here as a basis for description. And this does not mean naive falsificationism as long as there are attempts to explain away putative counterexamples by describing them in a different way. If the new description turns out to be better on independent grounds, both theory and description will have gained.

An open-ended project like the one outlined above requires of course the collaborative efforts of many linguists. It is therefore conceived of as an internet database system that after the test phase will be opened to the inter-
Realizing Humboldt’s dream

ested linguistic community. It is hoped that this outline will inspire many colleagues to take part in this great enterprise.

Notes

1. I am grateful to Nick Evans for offering me the opportunity to write about my favorite project in such an appropriate context. I am also deeply indebted both to him and to Alan Dench for a host of valuable questions and comments which were of great help to me in revising this paper and to Andrea Schalley for carefully checking and commenting on the revised version. All remaining errors and mistakes are therefore not their fault, but exclusively mine.

2. “...[man] hat aber auch diese [sc. die allgemeine Grammatik] ... nie als eine allgemein vergleichende Grammatik behandelt ... Zu noch größerem Unglück hat man fast überall daran [sc. an die zusammengetragenen Materialien] Urteile geknüpft, denen man es nur zu sehr ansieht, daß es ihnen an der sicheren Grundlage gehörig aufgestellter leitender Ideen fehlt.” (Humboldt 1996: 14)

3. For recent documentations of the software architecture and the linguistic categories see Nickles (2001) and Peterson (2002), respectively. Their work as well as that of the other project members was supported by grants Za 111/7-5 and Za 111/7-6 from the DFG (German Research Association) to the author, which are hereby gratefully acknowledged. The location of the project’s homepage is: http://www.crg.lmu.de

4. This is not meant to imply that the problem of version tracking and maintaining consistency is trivial.

5. Alan Dench points out (p.c.) that reaching descriptive adequacy can be rather challenging. One might add that any increase in the quality of a phenomenon’s description corresponds to a decrease in the difficulty of its explanation.

6. This does not mean that there is no way to correct assumptions that are built into the very description of the data: They cannot be literally falsified but if they turn out to yield very contrived descriptions this may lead to their revision.

7. The independence of the degree of complexity of the perceivable form from that of the entire sign (simple signs tend to have complex, multisegmental forms) is often referred to as double articulation.

8. This proviso serves to exclude identification, which is also sometimes subsumed under predication. Sortal signs can be used as the core of identificational predications. In the English sentence “It’s me,” the topical phenomenon is identified with the speaker and both are only sortally characterized. (I am indebted to Leonard Talmy (p.c.) for the nice example.)

9. Note that the restrictive predicative characterization of an entity in an act of reference as in ‘This woman’ is different from predicking the same concept
of a referent as in ‘This is a woman’, as can be seen in cases like ‘This woman is not a woman’, where the predication wins over the restriction if the sentence is not interpreted as contradictory.

10. The usefulness of Brentano’s distinction between the thetic and the categori-
cal judgement has been discovered for linguistics by Kuroda (1972), cf. also Sasse (1987), Ladusaw (1994).

11. This assumption is further developed and spelled out in Zaefferer (2001).

12. This involves a slight shift in reference: Illocutions are actions and even with assertions one would not say: ‘What you have done is true’, but rather: ‘What you have said is true’.

13. Imagine someone saying ‘You guys be quiet!’ to a silent audience.


17. My deep gratitude goes to Wallace Chafe who not only let me use his material but also willingly answered my email questions on the syllabification and interpretation of this example.

18. A few remarks on the notation (abbreviations are spelled out in the appendix): Levels +2 and +1: Ambisyllabic segments are underscored; level -1: Small caps gloss grammatical components, semantic components are represented in italics and their categories in small caps italics, the factors of polyfactorial morphemes are separated by dots; level -4: a is an action variable followed in square brackets by conditions on its illocution type and propositional content; s, is the object situation of the proposition, followed by its type; p is a variable for an elementary proposition type followed in square brackets by conditions on its referent and predicate; x, y, z are individual variables followed in square brackets by predicates they instantiate; e, e1, e2 are eventity (cf. fn. 19) variables followed in square brackets by partial characterizations of their aspectuality, type, and participants. ‘exponent’ designates the unspecified participant role. The variables are intended to be interpreted like discourse referents, i.e. as existentially bound.

19. Note that the orthographic level is not part of the example proper, it has been added to the phonological and morphophonological ones only for the sake of convenience.

20. ‘Eventity’ is a term meaning ‘event or similar entity’ and is preferred by the author over ‘eventuality’, which is used in linguistic circles with the same meaning, but has the disadvantage of already being lexicalized in the sense of ‘possibility’.

21. Among others, the Interlinear Representation Format is intended to be a framework for competing descriptions of the same data.
Abbreviations

3     third person
ACC    accusative
ACT    active
CAUS   causative
CONT   continuative
DAT    dative
DECL   declarative
DEF    definite
DIR.OBJ direct object
EXT    extension
ILL.TYPE illocution type
IND    indicative
LOC    locational
M      masculine
N      neuter
NOM    nominative
P      pre- or postposition
PAT    patient
PERSP   personal pronoun
POSSP   possessive pronoun
PRED   predicate
PRES   present tense
PRFX   prefix
PRTCL   particle
REP    repetitive
S      singular
S$_T$-ACCESSIBLE accessible in the interpretation situation on $s_T$
S$_I$ object situation
S$_P$-SAP speech act participant in the production situation on $s_P$
SBJ    subject
STAT   stative
SUFF   suffix
V      verb
References

Broschart, Jürgen

Comrie, Bernard, and Norval Smith

Comrie, Bernard, William Croft, Christian Lehmann, and Dietmar Zaefferer

Comrie, Bernard

Dixon, R.M.W.
1997 *The Rise and Fall of Languages*. Cambridge: Cambridge University Press.

Dryer, Matthew
2006 Descriptive theories, explanatory theories, and Basic Linguistic Theory, [this volume.]

Humboldt, Wilhelm von

Kuroda, Sige-Yuki

Ladusaw, William A.

Mosel, Ulrike
2006 Grammaticography: The art and craft of writing grammars, [this volume.]

Nickles, Matthias
Realizing Humboldt’s dream

---

Peterson, John
[http://www.cis.uni-muenchen.de/CISPublikationen.html]

Prillwitz, Siegmund et al

Sasse, Hans-Jürgen

Talmy, Leonard

Zaefferer, Dietmar
2003 A unified representation format for spoken and sign language texts.
[http://emeld.org/workshop/2003/papers03.html]
2004 Competition is Good for Descriptions: For a Consensus on Dissenting Entries as Desideratum for Linguistic Database Design.

Zaefferer, Dietmar (ed.)