CLIPP

Christiani Lehmanni inedita, publicanda, publicata

titulus
Language documentation. A program
huius textus situs retis mundialis
http://www.uni-erfurt.de/sprachwissenschaft/personal/lehmann/CL_Publ/Language_documentation.pdf
dies manuscripti postremum modificati
20.02.2001
occasio orationis habitae
DGfS Summer school 'Linguistic typology', 30.08. – 11.09.1998, Mainz
volumen publicationem continens
Bisang, Walter (ed.), *Aspects of typology and universals*. Berlin: Akademie Verlag (Studia Typologica, 1)
annus publicationis
2001
paginae
83-97

Language documentation
A program

Christian Lehmann

Universität Erfurt

20.02.01

Abstract

A language is a human skill and its products, comparable to some handcraft like weaving and its products. To document it means to preserve show pieces of it and to provide them with information which enable one to appreciate the skill and to learn (and thus, if necessary, revive) it to some extent.

Documentation of a language is an activity (and, derivatively, its result) that gathers, processes and exhibits a sample of data of the language that is representative of its linguistic structure and gives a fair impression of how and for what purposes the language is used. Its aim is to represent the language for those who do not have access to the language itself. **Description of a language** is an activity (and, derivatively, its result) that formulates, in the most general way possible, the patterns underlying the linguistic data. Its aim is to make the user of the description understand the way the language works.

While documentation in this sense has been done for other human skills, for instance in relevant museums, nothing of the sort has ever been suggested for languages. Therefore, many important questions have yet to be answered before one can even think of beginning. A methodologically interesting question is the following: Given this distinction between documentation and description, what kind and amount of information does the documentation have to add to the raw data so that a linguist should be enabled to come up with a description of the language on the basis of its documentation? From the answer to this and related questions, some proposals for a program of language documentation will be derived.[1]

---

[1] Versions of this contribution were presented at the Workshop on the 'Best Record' of the Max-Planck-Institute of Psycholinguistics at Nijmegen, 9 – 10 October 1995, and at the DGfS Summer School on Language Typology at the University of Mainz, 30 August – 11 September 1998. I thank the participants and Denny Moore for helpful discussion.

1. Data collection in linguistics

1.1 Data of living vs. corpus languages

As in any empirical science, a basic step in any linguistic work devoted to a language is the collection of data. To the extent that there has been a methodological tradition in linguistics, there is no unified role that data collection would play in it.[2] First of all, a basic distinction must be made between corpus languages and living languages. A philology or linguistic description devoted to a **corpus language** has to base all of its achievements on a corpus of data which is either finite or at any rate not extendable at will. Inductive generalizations are limited by the corpus. Even deductive generalizations must find evidence in the available data, or else they will be speculations. Consequently, the datum acquires a high value. Much of the scientific work is devoted to digging out, ordering, purging, preparing, interpreting, archiving, preserving, representing and transmitting linguistic data. This activity has a high prestige in the philologies. In Classical Philology, for instance, the best scholars have invested considerable portions of their lifetime in the edition of the texts. In such disciplines, linguists are either themselves engaged in this kind of activity or depend entirely on its results.

Things are completely different for a **living language**. There may be a corpus of data, too; but it is extendable at will. Consequently, descriptive work often has more data at its disposal than the investigator wants to look at. If a datum of the desired kind is not at hand, the investigator has ways of producing it. As a consequence, the value of language data in such a discipline decreases to a minimum. Data collection seems a completely ancillary and trivial activity in the linguistic description of a living language.

1.2 Data collection in epistemology

While the situation of the data themselves may be considered as something that a science is confronted with, but cannot change, there are also reasons in epistemology, or perhaps one should rather say, in the ideology of science, which contribute to the low prestige of data collection even in empirical disciplines. The highest goal of any empirical science is a theory. Worthwhile insights into its subject matter are of necessity general. A particular utterance is of no interest in itself, but only insofar as it instantiates a syntactic pattern of the language. The syntactic description of a particular language is of limited value in itself; it acquires higher value insofar as it represents a linguistic type. A linguistic type in itself is only part of a universal theory of language, which is what linguistics ultimately aims at. In such an ideology of scientific work, even descriptive linguistic work has a limited prestige because it is subordinate to explanatory theoretical work. Collection of data, however, has no prestige at all. It is the lowest step in a hierarchy, a step that one seeks to leave behind as soon as possible and that one skips more often than not.

I have called this epistemology of linguistics an ideology, thus implying that there are metaphysical convictions behind it. One such conviction is that language is an object of nature

---

[2] Cf. Lehmann 1996 for some discussion.

such as the objects of biology or chemistry. This means that the object has no individuality, that it is of scientific interest not as an individual, but as a member of a species. Considering linguistics as an empirical science very often implies this conception of the linguistic object.

The conception is, however, not justifiable on scientific grounds. Every language is an activity grounded in the historical situation of its speech community and also a factor in its history. Insofar, it is a historical object, and linguistics is a historical discipline (cf. Coseriu 1979, 1980[H]). A historical object has an individuality and therefore commands an interest for its own sake.[3] To the extent that each language is unlike any other language, its description is not transcendable by generalizations or universal theories. Nor is this a defect on account of which it would loose in value. This would appear so only from the vantage point of a certain ideology of science. In the perspective of an educated and intellectually active human being, the historical object has so much more value just because of its uniqueness.

The epistemological conception of language as a natural or as a historical object does not only shape the kind of interest that we take in it, but even the presupposed data situation. The utterances that constitute the basic data of a living language are infinite and replaceable only if we abstract from their individuality. Utterances, however, are historical objects. They represent a certain diachronic, diatopic, diastratic and diaphasic variety of a language (cf. Coseriu 1980[P]). Insofar, scientific data which embody utterances even of a living language are not replaceable at will. The possibility to replicate a given linguistic datum depends on the viability of the language variety that it is taken from and on the extent to which scientists have access to it. Data from a language which is the native idiom of many linguists can be replicated much more easily than data from a language spoken at the brim of civilization. Data from an extinct language cannot be replaced at all.

The outcome of these considerations is: Linguistic data which are not easily replicable demand the same care as the text corpus transmitted by generations of philologists. The corpus linguistics typical of the languages of antiquity, on the one hand, and the field linguistics typically devoted to languages of the third world which are far removed from traditional European culture, are two activities within the broad field of linguistics which are often considered as entirely unrelated. Surprisingly, they now find themselves sitting in the same boat, united by a common concern for irreplaceable data which are precious because they represent a historically unique facet of human life.

In the scientific ideology criticized here, linguistic data collection serves the sole purpose of rendering linguistic description possible. Linguistic descriptions – i.e., mostly, grammars and lexicons – derive their scientific status not from the quality of the data they present, but exclusively from their "descriptive and explanatory adequacy".[4] Therefore, they do not present primary data for their own sake, but only to the extent that this is inevitable in order to plausibilize the analysis or render it more easily intelligible for human readers. Data are very often

---

[3] "Die einzelnen Sprachen sind nicht als Gattungen, sondern als Individuen verschieden; ihr Charakter ist kein Gattungscharakter, sondern ein individueller." (Humboldt 1827-29:289)

[4] Recall the three-step hierarchy of observational, descriptive and explanatory adequacy in Chomsky 1965, ch. 1.4.

handled with utter negligence in linguistic publications. Similarly, in those circles which regard linguistics as a natural science, it is not a valid criticism in a book review to find fault with the data, since what is of scientific interest is not the data, but the analysis.

To do away with this ideology implies the rejection of the view of data collection as an inferior and ancillary prescientific activity. It does not, of course, mean to deny that data collection is a methodological step presupposed by linguistic description. Quite on the contrary, it is a necessary precondition for the latter. On the other hand, if we claim that linguistic data have a dignity of their own, do we mean that data collection is a self-sufficient goal? This is certainly not the case. There are, in linguistics just as in other sciences, data cemeteries, large amounts of data assembled by people who thought they served a purpose in themselves and not used by other people, who thereby prove the former people wrong. The immediate lesson from this is that linguistic documentation does not reduce to linguistic data collection.

2. Documentation in linguistics

2.1 The concept of documentation

The idea and the term of *language documentation* are novel in linguistics.[5] Linguists have not thought of their activity as a documentary one. While there are a variety of things that one can do to and with a language, such as describing or explaining it, documenting it is not an approach to language that would be familiar in linguistics. It seems therefore appropriate to start with some conceptual clarification.

In the past decades, we have come to associate with the concept of documentation the whole field of information storage and retrieval, of library resources, archiving methods and computer databases. Many a linguist has a more or less extensive linguistic documentation project which is essentially an annotated bibliographic database. Documentation in this sense means something like the systematic collection and representation of records of key information on some class of objects by which one can either get more information on an object or access the object itself.

For present purposes, this sense of the term, although certainly relevant, is not of central interest. Instead, a bit of etymology will be helpful here. The Latin word *documentum* is formed on the derivational pattern in T1.

T1. *Latin derivations in* -mentum

V-mentum	`thing for V-ing, thing that serves to V'
ornamentum	`thing for adorning'
monumentum	`thing to admonish'
documentum	`thing for teaching'

[5] General linguistics here is a bit behind the development in the philologies, in anthropology and literary studies. See Heissig & Schott (eds.) 1998 for a recent survey.

Accordingly, the literal sense of *documentum* is `thing that serves to teach'. It also means `example, proof'. By its etymology, *document* is thus a functional concept which evokes the question: who teaches what to whom? We will come back to such questions below. At the moment, it suffices to note that documenting a language involves producing and storing such records as may serve to teach one about the language.

The observant reader will have noted that I might have said "... such records as may serve to teach the language." Although this may be an important mediate goal of language documentation, it is not its direct purpose. To teach a language involves a presentation of linguistic material which proceeds from simple and everyday phrases to complex and peripheral constructions and which helps the learner in assimilating the material and in stepwise building a competence. In order to achieve this goal, the user has to get involved in speaking the language himself. This idea of learning by doing is irrelevant to documentation. The **primary purpose of language documentation** is to represent the language for those who do not have direct access to the language itself. What is essential here is not the motivation of the user. He may be a linguist or a layman, he may wish to find out about the way the language works, may be curious about its peculiarities[6] or may want to apply universal theories to it, or he may wish to learn it. A language documentation may be put to all these uses, but they are all tributary to the primary use of just representing the language itself. The essential issue here is, consequently, representativeness and accessibility.[7]

2.2 Documentation of a skill

Language is among those activities in which human beings are creative both as individuals and as members of a historical society. Insofar, it is comparable to such skills as house-building, music, pottery (cf. Silverman & Parezo (eds.) 1995). By pursuing the analogy, we may note that t e a c h i n g somebody the traditional pottery of a society normally involves making him sit down at the potter's wheel and work the clay. D o c u m e n t i n g the pottery, on the other hand, obviously starts from the orderly presentation of selected show pieces of different kinds. In addition, there may be various kinds of information. There may be photos or a video showing a potter in the various production phases. There may be explanations about the steps to take, from the quality of the clay and where it is found up to the temperature and duration of the burning. There may be information on the uses that the ready pieces of pottery are put to or on potters and their life. All these kinds of information would appear, e.g., in a good pottery exhibit of an ethnological museum.[8] It is easy to see that while such a documentation is not a course in pottery,

---

[6] Thus, an important specific purpose of language documentation is to serve as a record of the past and as an element of ethnic identity for future members of the community that has lost its identity as a speech community but which still recalls that their ancestors had a language of their own.

[7] In putting it this way, I do not, of course, exclude the possibility of combining the efforts for documenting a language with those for preserving it. It would be quite conceivable to use the documentation of a language as a chrestomathy (see §?). On the other hand, elaborating teaching materials of a language is obviously a different task than documenting it for posterity.

[8] Cf. Lehmann 1992 on the conception of a language museum.

it might play a fundamental role in designing one. Imagine a situation where the traditional pottery of a society has become extinct. If there is a comprehensive documentation of the kind indicated, then a skilled craftsman should be able to learn and thus revive that pottery from working through the documentation. Also, a ceramologist should be able, by relying solely on the documentation, to derive a scientific description of this variety of pottery.

We may now abstract from the example. The documentation of a human skill is goal-oriented in the sense that it pays primary attention to the things produced by that skill. If the products are volatile, as they basically are in the case of music and language, then they first have to be recorded on an enduring medium. That is, documentation of a language does not involve life representations of native speakers, just as documentation of a music style does not necessitate the presence of musicians. What we demand is that the show pieces be of high quality and representative of the variation.

In some cases, a good presentation of ready products may be sufficient for an experienced craftsman or even a layman to recognize how these things are made. For instance, by studying a traditional wooden Maya house, it should be possible – given sufficient dexterity – to build one oneself. In other cases, this would clearly not suffice. For instance, even a trained musicologist would be unable to find out how a certain variety of Papuan ethnomusic is made if he could just listen to a high-fidelity record of some piece, but had never seen any of the instruments. Similarly, the comprehensive documentation of a Hottentot language would include the presentation of a sagittal cut through the mouth which demonstrates the production of a click.

In some areas, the user of a documentation will still be unable to understand what is going on if he were just presented with products and movies of the production process. Imagine the documentation of a religious ceremony. We may see the priest signing the cross over a person. If we are not familiar with the culture, we will just observe him making a hand movement without knowing what he was really doing. What we need is an interpretation of the act. While this is obvious for all symbolic actions, it is even true for many other actions whose purpose is not evident from directly observable effects. As an example, think of a documentation of traditional healing methods. From such examples, we may generalize that the documentation of a skill does not reduce to the presentation of primary data. The user of the documentation must be put in a position of appreciating what is being done. This requires the addition of interpretive information which is not inherent in the primary data, but generated by the author of the documentation.

One may regret this necessity in the interest of objectivity. However, the mere collection and reproduction of raw data is not a scientific activity at all. Moreover, objectivity in documentation is an illusion, anyway. Since the documentation does not repeat the documented reality itself, but only represents a sample of it, there is necessarily a process of selection, which in itself is not objective and which, in fact, can be highly tendentious. A well-known example is the presentation of hieroglyphic inscriptions of the ancient Mayas by early epigraphers. The calendar was the only thing in these inscriptions that they had deciphered. Therefore inscriptions with extensive dating abounded in their documentations, and these were the only parts that were provided by a translation. As a result, americanists around the middle of our century believed that the whole spiritual life of the Mayas revolved around measuring the time and observing the calendar.

Human skills are embedded in the life of the society. Some aspects of traditional pottery will be understood only if we know about the position of the potter in the society. Similarly, certain passages in the documented texts and certain words will be understood only if they are provided with information about the reality being denoted. Although this is a trivial truth, its appreciation in linguistics has varied. In German linguistics around the second world war, there was a movement called `Wörter und Sachen' (words and things) which seeked to elucidate word meanings by factual knowledge about the designated objects of traditional culture and vice versa. American structural linguistics in its traditional combination with ethnology has cultivated this line of research, too. The two strands mentioned make us see that the endeavor to fully document a language transcends the boundaries of linguistics. A language documentation would not be feasible if ethnographic information were systematically excluded from it.

3. The relation between documentation and description

Documentation of a language is an activity (and, derivatively, its result) that gathers, processes and exhibits a sample of data of the language that is representative of its linguistic structure and gives a fair impression of how and for what purposes the language is used. Its aim is to represent the language for those who do not have access to the language itself. **Description of a language** is an activity (and, derivatively, its result) that formulates, in the most general way possible, the patterns underlying the linguistic data. Its aim is to enable the user of the description to compare this language to other languages.

In the clearest cases, the difference between documentation and description is a difference in the logical level. The documentation consists of specimina or representations of its object, while the description is a discourse on the object. If the object is a language, then the description bears a metalinguistic relation to the object, while the documentation remains on the object-language level. This relationship is shown in T2. However, as we shall see shortly, this is really a prototypical distinction. In practice, the purpose of a documentation is often not served sufficiently if no explanations are given (cf. §2.2). Explanations, however, belong strictly to the description. Moreover, in the cases of objects like languages, the documentation involves representations of the object; and a representation of an object is not easily assignable to the object-level or meta-level.

T2. *Documentation and description of a language*

<table>
<tr><td>description of language</td><td>meta-level</td></tr>
<tr><td>documentation of language</td><td rowspan="2">object-level</td></tr>
<tr><td>language</td></tr>
</table>

Other differences between documentation and description are derived from this basic one. There is a difference in generality in the sense that the documentation cares for the specific object, while the description seeks to generalize over it. And there is a corresponding difference in abstractness: the documentation is concrete, its components are easily perceived by the layman, while the description is abstract and made for the specialist.

It is beyond dispute that linguistic descriptions cannot be understood by laymen. Let us assume that this is a virtue, because otherwise it would be less clear what the specifically scientific about them is. There are, however, many linguistic descriptions – grammars and dictionaries – which are commonly regarded as sufficient pieces of their kind within the confines of the discipline. One might then ask: if we have such a description of a language, then what do we need a documentation for? In particular, if we have a generative grammar, then it can generate data for us, so there is no use in storing data at all.

In theory, documentation and description of a language are mutually independent. One should, in fact, document a language in such a way that future linguists can derive a description from it (cf. Himmelmann 1993); and one should describe a language in such a way that future linguists can produce data on the basis of it. In practice, however, these demands on the quality of documentations and descriptions exceed the capacity of human linguists. Already Wilhelm von Humboldt found occasion to write the afterwards oft-quoted passage on grammars available to him:

> Gerade das Höchste und Feinste läßt sich an jenen getrennten Elementen nicht erkennen und kann nur ... in der verbundenen Rede wahrgenommen oder geahndet werden. Nur sie muß man sich überhaupt in allen Untersuchungen, welche in die lebendige Wesenheit der Sprache eindringen sollen, immer als das Wahre und Erste denken. Das Zerschlagen in Wörter und Regeln ist nur ein totes Machwerk wissenschaftlicher Zergliederung. (Humboldt 1836:418f)[9]

Things have not changed much since Humboldt's times. Most of the available linguistic descriptions, among them some which are held in high esteem in specialist circles, do not allow one to produce a natural text based on them. This is not to say that such descriptions are without value. It merely means that they serve a different purpose. Therefore, while we may hope for ideal descriptions to come forth in the future, for the time being it is safer to produce a language documentation if one wants a lively representation of how the language really works.

Moreover, the dividing line between documentation and description is not sharp. As I said before, the documentation does not reduce to a body of raw data as produced by the speakers, but includes representations of the data, representations produced by the linguist, e.g. a phonetic transcription, an interlinear morphemic gloss, a translation. If this is so, then the documentation contains an analysis. It presupposes a description, and vice versa. For these reasons, it is neither possible nor advisable to separate the documentation from the description.

The description of a language does not reduce to an account of the language system, i.e. a phonology, a grammar and a lexicon. It also contains an account of the ethnographic, social, genetic and historical situation of the language (cf. Lehmann 1989). If a full account of the language system is already something that has been achieved for very few languages indeed, then a description in this comprehensive sense is truly boundless. I will come back to the problem of

---

[9] It is just the highest and finest which is not recognizable on the basis of those separated elements and can only be perceived or sensed in connected discourse. It is only this which we have to regard as the true and first instance in all investigations which are supposed to penetrate into the living essence of the language. The dismemberment in words and rules is only a dead artifice of scientific analysis.

limiting the task in §6. At this point, the parallelism between the documentation and the description should be noted. It was said at the end of §2 that the documentation is not confined to the formal linguistic aspects of the phenomena, but extends to ethnographic aspects. The same is, of course, true for the description. Although documentation and description of a language belong to different levels, they may be analogous as to their internal systematic structure.

4. Representation and interpretation of linguistic data

For a living language, linguistic raw data are video films or, in default of these, audio recordings of communicative events. They have to be represented at various linguistic levels. The professional linguist distinguishes a large number of levels, among them the phonetic, phonological, morphological and syntactic levels. In addition, there are two kinds of translation, an interlinear morphemic gloss and a free translation. However, the documentation of a language does not require representations to be this diversified. On the one hand, this would presuppose phonological, grammatical and semantic analyses, which would require to postpone the documentation until after the completion of the description. This is undesirable in view of the many languages which are in urgent need of documentation. On the other hand, the documentation is not made exclusively for the professional. If the layman can draw profit from a documentation without a syntactic analysis, then the linguist should be able to do so, too.

With the acoustic rendering of the original utterance, the user has immediate access to the significans of the language sign. The next thing he wants is an understanding of the document, i.e., he needs access to the significatum. This can be represented in form of a free translation. With this, the user can grasp the sense of the text. So far, the representations only account for the particular utterance. They remain at the level of *la parole*. *La langue* becomes relevant only as a vehicle, just as in everyday life. In linguistic documentation, the language system is paid attention to for its own sake. In addition to the sheer sound and the sheer sense of the text, we want to see its linguistic structure represented. At the same time, we want to keep the documentation as free as possible of descriptive theory, in the spirit of division of labor between documentation and description and because documentation can, in practice, not always rely on an available description and because the documentation should be usable by laymen. The scientifically responsible way of representing the structure of an utterance with the lowest possible degree of sophistication is to identify the morphs and match each with its meaning. Thus, the elementary representation of the structure of the significans is an allomorphic representation, and the elementary represention of the structure of the significatum is an interlinear morphemic gloss. Accordingly, T3 includes three levels of linguistic representation above the level of the raw data.

T3. *Representation of primary linguistic data*

linguistic representation	free translation	parole
	interlinear morphemic gloss	langue
	allomorphic representation	
raw data	video/audio recording	parole

Since people without linguistic training tend to find morphemes and interlinear glosses too abstract, it would seem desirable to produce an alternate representation for them in which the allomorphic representation reduces to an orthographic one and the interlinear morphemic gloss reduces to a word-for-word translation.[10] Moreover, a good documentation has to solve the technical problem of presenting the three linguistic representations simultaneously with the acoustic rendering.

Let us briefly come back to the notion that one of the purposes of a language documentation is to serve as the basis for future descriptions of the language. To my knowledge, there are no investigations of the question of what quantity and quality of data of a language a trained linguist needs in order to come up with a description of it. There is good evidence from professional experience that bare raw data with no additional information whatsoever are insufficient. For instance, all cases of successful script decipherment involved some kind of historical or archaeological information or even a bilingue in addition to the texts themselves. Wherever such information is not available, as in the case of the Indus Valley script, decipherment fails. Similarly, I once participated in an attempt to do a linguistic analysis of a tape recording of an unknown language. It failed essentially, apart from the first sentence, which we figured out on the basis of the external information that it was a fairy tale and the subsequent hypothesis that the sentence probably meant "Once upon a time, there was an X." With video recordings, conditions for linguistic analysis might be more favorable. However, the purpose of a language documentation is not to serve as a riddle for smart linguists, but to allow insight into the way the language works. The easiest way of achieving this is obviously to identify the meaningful elements and indicate their meaning.

5. Criteria for documentation

5.1 Quality

The criteria for the selection of texts that go into the documentation are essentially two: quality and representativeness. The quality of a text concerns both its content and its form. As for the former, the documentation of a language is part of the documentation of its culture. Therefore, texts whose content is important in the culture, for instance prayers or myths or instructions to the youth, are more highly valued than, say, routine everyday conversations.

[10] The latter part of the task is, alas, not trivial, as becomes particularly clear in the case of polysynthetic languages.

Quality of the form means linguistic correctness and aesthetic beauty of a text, as it results from the linguistic skill of its authors. In the philologies, most attention has always been devoted to literary texts. While these may be problematic from the point of view of representativeness, there is no doubt that they range high in quality. From this perspective, the presence of literary texts in general purpose language documentations and their prominent role in published linguistic descriptions is certainly justified, although not, of course, to the extent of excluding any other text genre.

For spoken texts, the quality of linguistic performance includes such aspects as textual cohesion, stylistic appropriateness, richness in system resources, correctness of constructions and, last not least, accuracy of articulation. The latter is, so to speak, the acoustic counterpart to calligraphy, which would have to be sought for if literary texts of a community with a long writing tradition were to be documented. Needless to mention, the linguist, too, bears part of the responsibility, namely for the technical quality of the recording.

The problem of the quality of the text is bound up with the issue of purism and the freedom of the linguist to edit the recorded text. The first thing to be said here is that the more the linguist cares for optimal production conditions, the less need will there arise for editing. Secondly, editing can only mean the production of further representations of the recorded text; under no circumstances must the original recording be changed. Finally, whenever possible, responsible editing should engage the authors of the text themselves. The decision on the form in which their text is to be transmitted to posteriority is first and foremost theirs.

5.2 Representativeness

The task to compose a representative text corpus of a language is, again, a novel one in linguistics.[11] Its traditional counterpart in the philologies is known by the name of chrestomathy. This is a collection of texts designed to facilitate the access to a language for someone who learns it as a second language. Traditionally, it is essentially an anthology of literary texts. I have already commented on the task of supporting the learner of a language in §2, and on the role of literary texts in §5.1.

Representativeness of a language documentation means that it adequately represents the purposes and ways the language is made **use** of in the speech community and the structural properties and possibilities of the linguistic **system**. As for the latter, there is no direct way of guaranteeing representativeness. For one thing, the presupposition is that documentation of a language should be possible without prior analysis of the linguistic system. For another, since the virtues of the linguistic system are language-specific, there would by definition seem to be no universal method of bringing them out. It therefore seems best to directly approach the first requirement, viz. the representativeness of the sample as regards uses of the language. If this is defined on a genuinely linguistic basis, then it should also, derivatively, provide for structural representativeness.

This purely linguistic basis for the composition of a representative text corpus must be oriented by parameters which are universal in all languages. These are the parameters which constitute the speech situation. The essential components of the speech situation are the speech act participants,

[11] See Lenk 1996 for some methodological considerations.

the context, the communicative task, the topic talked about, the code, the channel and the message (cf. Jakobson 1960). Each of them constitutes a parameter of variation. The parameters together define a multidimensional space, in which text genres can be placed. T4 gives an overview of the variation generated by these parameters and of the ways they may be used for the characterization of text genres.

Some comments are in order here. First, not all of the components of the speech situation are made use of in T4. The message does not appear because this is just the dependent variable whose full range of variation we hope to obtain by varying the other parameters. The code does not appear because it is here equated with the language system, and this we want to keep constant. Again, it is assumed that code variation below the level of the language system, such as choice of different sociolects or registers, will follow as a dependent variable if we vary the other parameters, for instance the social group of the speech act participants or the formality of the speech situation.[12] The same goes for code switching. This is, of course, not excluded by the premise of focussing on one language system. It, too, should come out as a consequence of varying some of the other parameters. Likewise, diachronic variation (to the extent that it appears at the synchronic level) should be obtainable by suitable combinations of the parameters of T4. For instance, narrative monologs typically conserve older strata of a language, while creative dialogs are typically progressive in the use of linguistic resources.

[12] I assume, along the lines of Kabatek 1999, that the parameter of distance vs. proximity, which is postulated as the central universal parameter of linguistic variation in Koch & Oesterreicher 1990:14, can be reduced to other dimensions such as those mentioned.

T4. *Components of speech situation and text genres*

1. Speech act participants (speaker, hearer, bystander)
Nature (supernatural vs. human being vs. animal vs. none; monolog, dialog, palaver ...).
Social group (sex, age, social status, profession, ethnic affiliation ...).
Roles: symmetric vs. asymmetric (kin, chief vs. citizen ...), intimate vs. stranger.

2. Context of speech act (situation)
Place (church, mill, pub ...).
Time (daytime vs. night ...).
Formality: distance vs. proximity (friendly encounter, work, ritual ...).
Real-life embedding (game, drama ...).

3. Task
Illocution:
narrative (myth, proverb, joke, riddle ...);
instructive (working routine, game instruction ...);
discursive (political/forensic speech, sermon, blessing, curse ...);
interrogative (examination, interrogation ...);
poetic (poem, song ...).
Conventionality:
ritualized (baptism, courtship ...);
conventional (greeting and leave, route directions, official address ...);
creative (event report, dispute, poem ...).

4. Topic
(Traditional/modern; work/leisure, past/future, family/village, mythical figures/persons /animals/plants ...).

5. Channel
Medium: oral vs. written.
Directness: face-to-face vs. technical transmission.

Given these assumptions, a representative text corpus of a language may be defined as one which exhibits sufficient variation on each of the parameters of T4. This notion may be refined by spelling out the way in which the parameters cross-classify and by characterizing some parameters or values as more central than others. We will come back to this in the last section.

The two essential criteria of quality and representativeness are logically independent and therefore not seldom in conflict, as already observed in §5.1. Everyday utterances which are typical of informal encounters are certainly representative of the ways of communication in a society, but they may be of low quality with respect to cultural importance of their content or accuracy of articulation. If the language approaches extinction, most of the members of the dissolving speech community will be semi-speakers, who produce low quality in terms of richness in linguistic structure. The proper choice under such conditions obviously depends on the purpose which the documentation is to serve. A comparison with other fields of culture which preserve creations of

the past, such as literature or archaeology, would lead one to conjecture that high quality of the texts will prove most important for future generations.

This implies that the criteria proposed here for the documentation of a language also bear on the selection of languages to be documented in the first place. At first sight, it seems clear that priority should be given to endangered languages. There can, in fact, be no doubt that the documentation (and the description) of those few hundred languages which are not endangered, but which nevertheless currently absorbs 99% of all linguistic resources, is actually not urgent at all and can safely be postponed until all the languages now endangered have become extinct (cf. Lehmann 1998). On the other hand, it would be hasty to conclude that the priority of documenting a language is the higher the more the language is endangered. This would be so only if we had infinite means for language documentation at our disposal. As long as manpower and funds are limited, large quantities of languages will, despite all our efforts, die out before they have been sufficiently documented. If, under these circumstances, we decided to dedicate all available means to whatever language is, at the moment, most endangered, we would always be documenting moribund languages which only have semi-speakers left. On the one hand, the ratio of gains and losses is particularly bad for these languages because data gathering is exceedingly difficult and time-consuming. On the other hand, such a decision would lead us into conflict with our two documentation criteria. For a language which only has semi-speakers left, neither quality nor representativeness in documentation can be achieved, because those few speakers do not remember the language well enough, since they have not used it for most purposes for a long time.

The conclusion from this is that, as far as assignment of available manpower and funds is concerned, priority should be given, *ceteris paribus*, to those languages which are yet lively enough to allow at least for a minimum standard in quality and representativeness of documentation. If we have to accept that many languages will die out without documentation, then let them be those languages which we cannot decently document, anyhow. I know that this proposal will hurt several colleagues' feelings; but I know of no viable alternative.[13]

6. A program for language documentation

There are many languages which are in urgent need of documentation because they are becoming extinct. Professional experience tells us that it may take a linguist's lifetime or even more to fully describe a language. If a documentation can be the basis for future descriptions, and if a documentation can be completed considerably faster than a description, then documentation projects would be the first choice for many languages in the world today. However, since the concept of language documentation is novel in linguistics, we have no basis in experience for making a sensible estimate of the average duration of a language documentation project. Whoever has worked on the edition of a field-recorded text knows that this is very time-consuming. It would be easy to spoil the whole enterprise of language documentation from start by setting the

[13] As I am writing this (october 1999), the Volkswagen Foundation has launched a program of documentation of endangered languages based on just these principles.

standards too high. Therefore, we have to allow for different degrees of completeness in various dimensions.

The Prague school conception of **center and periphery** in language will be helpful here. While T4, if suitably specified, can serve to define dimensions on which maximum variation is desired, we can, at the same time, arrange the values of each parameter according to their degree of centrality in the communicative life of a speech community. For instance:

- On parameter 1, a constellation in which both the speaker and the hearer are adult human beings of the same social group, for instance a married couple, may be taken to be highly representative, while other constellations are increasingly marginal.
- On parameter 2, a real-life situation is more central than a play.
- On parameter 3, a conventional task such as giving route directions is more central than a creative task such as composing a poem.
- On parameter 4, oral face-to-face communication will be taken as a core case, while telephone conversations or letter writing are more at the periphery.

In this way, we will be able to define a gradience between core and periphery in the language. To spell this out is among the most urgent tasks of our discipline. A documentation project may then start at the core and proceed to the periphery. However generous or limited the resources of such a project are, such a procedure would seem to always guarantee the best documentation possible under the given circumstances.

References

Chomsky, Noam 1965, *Aspects of the theory of syntax.* Cambridge, Mass.: MIT Press (Special Technical Report No. 11).

Coseriu, Eugenio 1979, "Humanwissenschaften und Geschichte. Der Gesichtspunkt eines Linguisten." *Jahrbuch der Norwegischen Akademie der Wissenschaften*. Historisk-filosofisk klass, 10. november 1978; 3-15.

Coseriu, Eugenio 1980, "`Historische Sprache' und `Dialekt'." Göschel, Joachim & Ivi , Pavle & Kehr, Kurt (eds.) 1980, *Dialekt und Dialektologie. Ergebnisse des Internationalen Symposiums "Zur Theorie des Dialekts", Marburg/Lahn, 5.-10. September 1977*. Wiesbaden: Steiner (Zeitschrift für Dialektologie und Linguistik, Beihefte N.F. Nr. 26 der *Zeitschrift für Mundartforschung* [sic!]); 106-122.

Coseriu, Eugenio 1980, "Vom Primat der Geschichte." *Sprachwissenschaft* 5:125-145.

Heissig, Walther & Rüdiger Schott (eds.) 1998, *Die heutige Bedeutung oraler Traditionen. Ihre Archivierung, Publikation und Index-Erschließung.* Opladen: Westdeutscher Verlag (Abhandlungen der Nordrhein-Westfälischen Akademie der Wissenschaften, 102).

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

Humboldt, Wilhelm von 1827-29, "Über die Verschiedenheiten des menschlichen Sprachbaues." Humboldt 1963: 144-367.

Humboldt, Wilhelm von 1836, *Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluß auf die geistige Entwicklung des Menschengeschlechtes.* Berlin: Königl. Ak. Wiss.; in Kommission: Bonn etc.: F. Dümmler. Abgedr.: Humboldt 1963:368-756.

Humboldt, Wilhelm von 1963, *Schriften zur Sprachphilosophie*. [= *Werke in fünf Bänden*, hrsg. v. A. Flitner und K. Giel, Bd.III]. Darmstadt: Wissenschaftliche Buchgesellschaft. 4. Nachdruck 1972.

Jakobson, Roman 1960, "Closing statement: Linguistics and poetics." Sebeok, Thomas A. (ed.) 1960, *Style in language*. Cambridge, Mass.: MIT Press; New York & London : J. Wiley & Sons; 350-377.

Kabatek, Johannes 1999, L'oral et l'écrit – quelques aspects théoriques d'un «nouveau» paradigme dans le canon de la linguistique romane. Tübingen: unpubl. ms.

Koch, Peter & Oesterreicher, Wulf 1990, *Gesprochene Sprache in der Romania: Französisch, Italienisch, Spanisch.* Tübingen: M. Niemeyer (Romanistische Arbeitshefte, 31).

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

Lehmann, Christian 1992, "Das Sprachmuseum". *Linguistische Berichte* 142:477-494.

Lehmann, Christian 1996, "Dokumentacija jazykov, nakhodjaš ikhsja pod ugrozoj vymiranija. (Pervoo erednaja zada a lingvistiki)." *Voprosy Jazykoznanija* 1996/2:180-191.

Lehmann, Christian 1998, "Das große Sprachensterben." Fakultät für Linguistik und Literaturwissenschaft der Universität Bielefeld (ed.), *25 Jahre für eine neue Geisteswissenschaft.* Bielefeld: Aisthesis; 131-151.

Lenk, Elena 1996, Erstellung eines Textkorpus für eine Sprachdokumentation mit LDS. Bielefeld: Universität (AVG Arbeitspapier Nr. 13).

Silverman, Sydel & Parezo, Nancy J. (eds.) 1995, *Preserving the anthropological record.* New York: Wenner-Gren Foundation for Anthropological Research. 2. ed.