elpublishing

Language Documentation and Description

ISSN 1740-6234

This article appears in: *Language Documentation and Description, vol 5*. Editor: Peter K. Austin

On the representativeness of language documentations

FRANK SEIFART

Cite this article: Frank Seifart (2008). On the representativeness of language documentations. In Peter K. Austin (ed.) *Language Documentation and Description, vol 5*. London: SOAS. pp. 60-76

Link to this article: http://www.elpublishing.org/PID/063

This electronic version first published: July 2014

On the representativeness of language documentations

Frank Seifart

1. Introduction[1]

Speakers usually communicate in a language in all kinds of different constellations at different places and times about all sorts of things, provided that the language is not moribund. A very real, practical problem for someone wanting to document languages (in the sense of Himmelmann 1998, 2006) is thus where to point the camera and microphone and when. The cumulative result of these individual documentary recordings will be a corpus of primary data, the centre piece of a language documentation. It is desirable for this corpus to be representative of the language, the more so if it is endangered and most likely soon not be spoken anymore. The importance of representativeness is maybe most apparent when considering the opposite case: a misrepresentation of a language (and thus the speech community) by a heavily biased corpus. Imagine that at some later stage of history, the data available about our western-style civilization would be heavily biased towards (or even limited to) Grimm's fairy-tales (or soap operas, or theoretical physics).

This paper approaches the problem of representativeness of a language documentation by first discussing various criteria that may be used to select the events that are recorded and included in the documentation (section 3). Central to this problem are the possibilities and limitations of applying criteria that are based on a systematic classification of communicative event types and may therefore help to ensure the representativeness of a documentation in a theoretically grounded way (section 4). We then discuss the application of such criteria in a documentation project in the North West Amazon (section 5). The main conclusions from these discussions are that representativeness must be based on a careful analysis of culture-specific event types and that it is not possible to achieve representativeness for all kinds of communicative events to the same degree (section 6). Before entering into the main discussion, the relation of the central problem discussed here to the theoretical framework for language documentation is treated in the following section.

[1] Many thanks for useful comments are due to Nikolaus Himmelmann, Peter Austin, and Friederike Lüpke. All errors and shortcomings are mine.

Frank Seifart 2008. On the representativeness of language documentations. In Peter K. Austin (ed.) Language Documentation and Description, Vol 5, 60-76 London: SOAS.

2. The format and contents of language documentation

This paper argues that criteria for representative documentation should be an integral component of a theory of language documentation. Therefore, we briefly review the state of the art of documentary linguistics (as formulated by Himmelmann 1998; 2006) in order to show how the development of such criteria relates to it.

The format of language documentation is conceived by Himmelmann (2006: 21; see also 1998) as in Table 1. A key feature of this format is the clear separation of primary data from any descriptive or analytical statement about this data, which is placed in the 'apparatus' of the documentation. The primary data is organized into 'sessions', where one session is a recording of a some kind of communicative event, i.e. ideally displaying a unity of speaker(s), place and time (see further discussion below). The apparatus is divided into one section which contains documents related to individual sessions and one for the documentation as a whole. Specific proposals have been made for the various components within the apparatus such as the format of annotations (e.g. Lieb and Drude 2000; Schultze-Berndt 2006) or the format of a descriptive grammar as a component of a language documentation (Mosel 2006b). As an extended cataloguing device, the metadata includes all relevant information on the circumstances of the recording, such as the time, place, speakers, etc. The format of the metadata for language documentation has been standardized after lengthy discussion into the IMDI metadata set (see http://www.mpi.nl/IMDI), which is now widely used. Thus, one may say that documentary linguistics is quite advanced in defining, theoretically grounding and standardizing the format of a language documentation, both its overall structure and its individual components. In comparison to this, relatively little attention has been paid to the contents of the documentation, i.e. the way the primary data is selected and how this selection is structured, since Himmelmann (1998: 176ff.) (see also Lehmann 2001: 90ff.). This paper attempts to take a further step towards filling this gap.

Figure 1: Format of a language documentation

Primary data	**Apparatus**	
	Per session	**For documentation as a whole**
recordings/records of observable linguistic behaviour and metalinguistic knowledge	*Metadata* *Annotations* transcription translation further linguistic and ethnographic glossing and commentary	*Metadata* *General access resources* introduction orthographic conventions glossing conventions indices links to other resources ..... *Descriptive analysis* ethnography descriptive grammar dictionary

There is consensus about one very general aspect of the contents of language documentation, i.e., that its focus is not so much on the language system (which is the subject matter of descriptive linguistics) but on the use of language in its culture-specific context.[2] The units of the contents are thus instances of language use. We follow Himmelmann (1998: 168) in calling these units 'communicative events', a term taken from the ethnography of communication (see section 4, below). As mentioned above, a recording of such an event, which is integrated along with corresponding metadata and possibly annotation in a language documentation is called a 'session'. Sessions may be very different in nature and length, ranging from a recording of a single word (e.g. for phonetic analysis) to recording a lengthy ritual or festival. The corpus of primary data, organized into sessions, is the central component of a language documentation, around which all other components are organized. The role of the apparatus is mainly to allow access to the primary data, by, e.g. organizing the sessions in a hierarchical corpus structure and providing translations and further explanations and commentary.

3. Types of criteria

At his point, it is useful to be more precise about what is meant by 'representativeness' of language documentation. It means that the selection of events which make up the corpus of primary data allows someone who is not familiar with the language and speech community to gain an authentic picture of how the language was used at the time that the documentation was carried out. I should also note here that what I mean by representativeness also includes 'completeness' (or 'comprehensiveness' Himmelmann 1998: 176ff.). But rather than focusing on an upper limit (as these terms may suggest), I focus on the internal structure. As a starting point, we may distinguish three basic types of criteria that may be used for selecting events[3] in order to see how they may contribute to the representativeness of a language documentation. These criteria are referred to here as sampling methods and

[2] In fact, it is not possible to directly document the system of a language, but this is rather the result of linguistic analysis and description. Ideally, however, a description of the language system can be done on the basis of a documentation, in particular if elicitation sessions are included, e.g. on complete paradigms which are often missing even in large corpora.

[3] To be precise with respect to the meaning of 'selecting events' also: What we mean here is the decision whether to record and include an event (and, in practice, actively seek an occasion to record it) or not. Not meant here are related (and likewise difficult) decisions such as how to adjust the camera angle (i.e. what to include or exclude of the physical surroundings) and where to begin and end a recording (i.e. what to include or exclude of the temporal surroundings). See Widlok (2004) for discussion of some of these points.

their labels are borrowed freely from statistics. These methods are interrelated in various ways, as will be discussed below.

1. **Convenience sampling** refers to recordings that are done as the occasion arises. There is in fact no conscious selection procedure but what is recorded and what is not recorded is a matter of coincidence. Inclusion of data by this criterion is justified by the aim to make a language documentation as sizeable as possible, i.e. by the fact that any (or almost any) recording is valuable if the alternative is no recording at all.

2. **Externally motivated sampling** means selecting events according to requirements of users of the language documentation. One group of potential users are researchers. They may be interested in, e.g., recordings of conversation for a study of turn-taking, recordings of children for a language acquisition study or recordings of traditional festivals for an ethnographic study of ceremonial systems. Another important group of potential user are members of the speech community and their descendants. It may be in response to a specific request by them that, for instance, a traditional festival is recorded, which the festival organizer wishes to have a video recording of. What characterizes sampling by this method is that the criteria used for the selection of an event are not (at least not directly) derived from properties of the communicative event itself, but rather from an external scientific framework (e.g. conversation analysis, language acquisition), which requires certain kinds of data for its further development, or from an interest or personal taste of a member of the speech community. Inclusion of events in a documentation by externally motivated criteria is justified by the requirement to make a language documentation as useful as possible for the largest number of potential users as possible.

3. **Systematic sampling** refers in general to the inclusion of sufficient examples of each recurring type in the data. Given the focus on language use, what is meant here is the inclusion of examples of communicative event types. This method presupposes previous identification of these types. Such identification may be achieved through the systematization of the events along a number of parameters which describe their components. This method is thus ── at least in principle ── based on properties of the data itself (see further discussion in section 4).

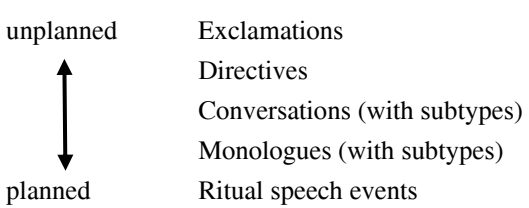
Both convenience sampling and externally motivated sampling must play an important role in any documentation project and both are justifiable and certainly produce useful data. On the one hand, coincidence (which underlies 'convenience sampling'), necessarily intervenes in any recording, at least in the sense that a recording can only be made at one certain place and one certain time, namely when the person documenting the language happens to be present. On the other hand, this person is likely to point the camera and microphone at an event that for some reason appears to be interesting for them. Assuming that this interest stems at least in part from their academic background (or any other acquired preference for the event, for that matter) the recording would also be a case of externally motivated sampling. Thus it is neither possible nor desirable to avoid convenience and externally motivated sampling all together. Also, neither of these methods are incompatible with systematic sampling. But it is with this latter method that it is theoretically possible to define representativeness in the sense envisaged here. For this reason, this method is discussed separately in the next section, where we also come back to the relation between externally motivated sampling and systematic sampling.

4. Systematic sampling of communicative events

As mentioned above, systematic sampling of communicative events is based on a classification of such events. The crucial question is thus to come up with such a classification, which reveals the recurrent types of communicative events that are used by the speech community. Various research traditions have approached the problem of systematically describing the variety of communicative events of a speech community using a number of parameters which define the communicative event types. In this section, we briefly review these approaches and discuss them in relation to documentary linguistics, before discussing their application in a case study in section 5, below.

An explicit proposal for systematic sampling is the 'spontaneity parameter' developed by Himmelmann (1998: 177ff.), based on Ochs' (1979) notion of 'plannedness' (Figure 2). Underlying this parameter is the finding that communicative events can be distinguished with respect to the kinds of linguistic structures that occur in them. Namely, the less spontaneous (and more carefully planned) a communicative event is, the more complex linguistic structures tend to occur in it. The spontaneity parameter does not define discrete types of events, but rather overlapping types along a continuum. The inclusion of examples of event types located at different points along this continuum thus helps to ensure a representative documentation of linguistic structures.

Figure 2: *The spontaneity parameter (Himmelmann 1998)*



Although primarily aimed at ensuring a representative documentation of diverse linguistic structures, this parameter distinguishes a broad variety of communicative event types in a more general sense, in particular if subtypes of the broad types such as conversation and monologue are identified. This parameter is also sufficiently general and operational to be applied to all speech communities. An important characteristic of this approach is that the variety of communicative event types is reduced to a single, powerful parameter (allowing for further distinction within broad types), from which other properties follow, such as complexity of linguistic structuring.

A different approach is to start out with a large number of specific parameters which describe in detail the components of individual communicative events. This is the method applied by the ethnography of communication. Taken together, these components are meant to provide a comprehensive description of culturally appropriate conduct in a communicative situation. An important feature of this approach is that the language (in the sense of linguistic systems) used in this situation is but one of a large number of characterizing parameters. This approach thus fits well in the broad perspective of language use in its cultural context that is the object of language documentation. The classic formulation of the ethnography of communication is found in Hymes (1971), who proposes the set of parameters givein in Figure 3 to comprehensively describe a communicative event, mnemonically organized according to the acronym SPEAKING (see also Saville-Troike 2003).[4]

[4] Lehmann (2001) proposes a somewhat similar list of variables for the description of components of speech and cites Jakobson (1960) in this context.

Figure 3: Components of speech in the ethnography of communication

S setting, scene

P participants (speaker, or sender; addressor; hearer, or receiver, or audience)

E ends (purposes, goals)

A act sequence (message form and content)

K key

I instrumentalities (channels and forms of speech, including code)

N norms of interaction and interpretation

G genres

These parameters set up a multidimensional space in which communicative events are identified. In theory a comprehensive description of possible communicative event types is thus obtained by keeping the values of some parameters constant while varying other parameters (this is envisaged by Lehmann 2001). For example, we may distinguish one event type with a male speaker from another with a female speaker (variation of parameter 'speaker'), all other variables being kept constant. As this example already shows, in practice the identification of event types is a much more analytical and creative process, which requires careful participant observation and a good knowledge of the language and culture. It will have to involve many more distinctions which are specific to the particular speech community and may be quite unexpected. To name just one example, kinship systems (as a sub-parameter which would have to be specified under the parameter 'participants') may differ drastically and bear on the event types in different ways, as in 'avoidance speech' in the presence of one's mother-in-law in some Australian languages (Dixon 1980: 58f.). The list of components given in Figure 3 should thus rather be understood (as explicitly stated by Hymes 1971: 53) as a heuristics to uncover the culture-specific communicative event types of a given speech community and the (also culture-specific) parameters that define them. For this reason, there is no need here to go in detail through all parameters suggested in this framework. Since the relevant parameters are not given *a priori*, the ones that are identified as the relevant ones and used in a language documentation should be made explicit, e.g. in a 'sampling methods' section in the introduction.

With respect to the applicability of the parameters suggested by the ethnography of communication, it should be made clear that these are not primarily intended as serving for a comprehensive classification of all event types in a given community. They are rather conceived for the description and analysis of individual types, often very specific ones, e.g. 'stylized sulking' among young Afro-Americans (Gilmore 1985). Consequently, studies within the ethnography of communication which do attempt a comprehensive description of event types within a speech community are very few (e.g. Sherzer 1983). Thus, while the ethnography of communication is a very helpful approach for developing criteria for representative language documentations, a lot of work within documentary linguistics remains to be done to obtain criteria that are appropriate for this specific aim. For instance, the parameter which identifies the language used ('code', which is one parameter under 'I – instrumentalities') is less relevant in a language documentation in the sense that it is usually being kept constant, at least in cases where a language documentation focuses on just one language (but see section 5, below). However, if one agrees that one of the aims of a documentation of an endangered language should be to document the process of language shift (Seifart 2000: 44), the 'language' parameter within a systematic sampling procedure should serve to ensure a representative documentation of the patterns of code switching and code mixing between the endangered and dominant language.

A major challenge to any attempt at a comprehensive classification and identification of communicative event types existing in a speech community is the identification of 'major' versus 'minor' types. The above-mentioned theoretically possible application of all parameters that were identified as relevant (and a theoretically infinite number of ever finer sub-distinctions) leads to a very large (potentially infinite) number of types. But some of these will be in some sense more different from each other than others and not all of them will be relevant for a representative documentation to the same degree, so criteria must be developed that allow us to distinguish major from minor types. The identification of major and minor types is also very important to make the sampling procedure efficient in the practice of a documentation project, since — given limited resources and time — not every, ever finer sub-distinction can be represented with examples.

It is clear from the discussion above that the identification of types which are sampled must be based on decisions that are made by a researcher and these thus affect the contents of a language documentation. These decisions are analytical and based on theoretical frameworks, thus they reflect properties of the data only through them. In fact, a systematic description of communicative event types is an abstraction of the data, obtained through analysis, not unlike a descriptive grammar in this respect. The systematic

sampling method as it is conceived here is thus similar to what has been called externally motivated sampling in the sense that both draw on theoretical frameworks. The difference between the two is not so much whether external frameworks play a role or not, but that the analytical (and thus maybe necessarily biased) perspective taken in 'systematic sampling' focuses the whole documentation, not individual aspects, with the explicit aim of representative documentation.

Before turning to a case study, it is useful to ask whether there are other important properties of events that may be relevant for systematic sampling. Firstly, it is easy to see how the spontaneity parameter can be applied in parallel to a comprehensive classification of events through parameters such as those mentioned above and independently help to ensure representativeness according to plannedness and thus complexity of linguistic structures.

Secondly, we may take frequency as another property of event types that can be taken into account for representativeness of documentation (see also Himmelmann 1998: 181). This would lead to the requirement of including a large amount of informal events. It is probably not reasonable to attempt a precise measure of the frequency of each event type (or informal events in general) and to match this measure in the corpus of primary data of a language documentation. However, a variety of informal communicative events in a documentation is necessary to give an authentic impression of the language as it is used in the context of culture-specific communication, even though recordings of these events are often not as highly valued as, e.g., elaborate formal events. One may add here that frequency as a sampling criterion should certainly also lead to the requirement to include at least roughly comparable portions of male and female speech.

5. Sampling songs and informal events in practice

In order to illustrate some of the points made above, we shall now briefly discuss some examples. The selection of events of two very different kinds will be discussed, which are located at distant points on the spontaneity parameter: firstly, songs performed at traditional festivals, i.e. a very formalized type of event, and secondly, scolding, i.e. a very informal type of event. The examples come from an ongoing documentation project, which focuses on the so-called 'People of the Centre', a multilingual cultural complex in the North West Amazon.[5] The People of the Center are a

[5] This project is carried out in collaboration with Doris Fagua, Jürg Gasché, and Nikolaus Himmelmann, among others, at the Insituto de Investigación de la Amazonia Peruana (Iquitos) and the Ruhr-Universität Bochum. The financial support of the Volkswagen Foundation, through the DoBeS program, is gratefully acknowledged.

culturally relatively uniform, but linguistically diverse group speaking seven mutually unintelligible languages which belong to three distinct linguistic groups (Arawak, Witotoan and Boran).[6] They fall into two cultural subgroups, distinguished by ceremonial systems. Their habitat is in South East Colombia and in neighboring areas in North East Peru.

A first point to be made here regards the selection of languages. For various — mostly practical — reasons, not all seven language can be documented to the same degree within the project. Two of the seven languages, Bora and Ocaina (Witotoan), are the subject of comprehensive documentations, consisting of fully annotated video recordings of a representative sample of each major type of communicative event, including ceremonial and ritual discourses, drum communication, as well as informal conversation. Another language, Resígaro has now only two native speakers left, with whom we record an as varied selection as possible of event types. From Witoto proper, we include old recordings of ritual discourses not practiced anymore, as well as new data. Taken together, this data set is a representative sample of the linguistic and cultural practices of the People of the Center in the sense that it includes comprehensive data from both cultural subgroups (Bora and Ocaina vs. Witoto, Resígaro and others), as well as data from the three linguistic families Witotoan, Boran, and Arawak.[7]

Among the unique cultural practices of the People of the Center are repertoires of thousands of songs performed at festivals. The singing performances at the festivals last for up to 20 hours, starting as early as noon, and never ending before dawn. The order in which the songs are sung is predetermined according to a complex scheme. There are about a dozen different types of festivals for each ethnolinguistic group of the People of the Center. Some of these types are specific to individual linguistic groups, while other festival types are celebrated by all groups. Each linguistic group has its own repertoire of thousands of songs for these occasions. Usually, at least two different language groups perform at one festival, alternating their performances within the predefined scheme.

We will first focus on one song type, which is the first song at the festival which is celebrated on the occasion of the inauguration of a new roundhouse (see Image 1 for a photograph of a group of Ocainas performing this song type at a Bora roundhouse). The parameters provided by the ethnography of

[6] Whether Boran and Witotoan belong to the same genetic group, as argued by Ashmann (1993), is still under debate (Kaufman 1994; Seifart in press). If they do, they are two very distantly related branches.

[7] Note, however, that convenience — or rather inconvenience — also plays a major role in this selection, namely the inaccessibility of some speakers in areas of South East Colombia which are occupied by guerrilla forces.

communication roughly describe this event type as follows: The **setting** and **scene** is the central part of the roundhouse of the festival organizer; the main **participants** are the invited group of singers and the organizer of the festival and his associates for whom the song is performed; the main **ends** of the event are to inaugurate the festival in a cheerful way and at the same time to criticize the festival organizer; the **act sequence** (message form and content) is a song which is composed according to a scheme which is quite strict both with respect to its musical characteristics (harmony and rhythm) and linguistic characteristics (preset phrases); the **key** in which the song is performed, i.e. in which the criticism of the festival organizer is conveyed, may be called mocking; the **instrumentalities** (channels and forms of speech, including code) are singing in the respective language; the **norms of interaction and interpretation** are complex and include that a group of men dance in one row and lead the singing, accompanied by a group of women that dance in a parallel row and accompany the singing in higher pitch voice; the **genre** that this song type belongs to may be called group-song, which is opposed to other song types which are performed individually.

Image 1: First song at the inauguration of a new roundhouse

First note that in order to fully describe this event type and, crucially, to differentiate it from other event types, i.e. other songs performed at the same festival type or songs of other festival types, it is necessary to introduce many distinctions within the general scheme provided by the ethnography of communication, such as the precise location of the singers, they way in which the song is composed, and the culture-specific festival types themselves. This shows how a general 'grid' set up by parameters such as those proposed by the ethnography of communication is in fact no more than a very general guideline indeed and how the identification and differentiation of individual event types is only possible through a very careful ethnographic and linguistic

analysis. That is, this technique is far from being a automatic, mechanical procedure.

In the case of festival songs of the People of the Centre, it is feasible to achieve a comprehensive classification of all types of songs, which is, by the way, much helped by the speakers' explicit knowledge of the classification of song types, many of which have names in the language. It yields about six different general types of songs for a festival like that of the inauguration of a new house. Among these are the first song (which is repeated once about 7 PM), songs sung in the daytime, a 'closing song' for the songs sung in the daytime, which is sung after every round of about four to five songs, songs sung at night from about 7 PM to 3 AM, a closing song for the songs sung at night, and a set of final songs, sung from about 3AM to 5 AM. A representative documentation of the songs of the festival for the inauguration of a new house thus proceeds by including sufficient examples of each of these types.

It is obviously desirable to document such a festival in its entirety, i.e. make a recording from beginning to end, in order to be able to appreciate the performance of the songs in their order and context. However, this is not always possible — and less so for all festival types of all language groups — given the limited resources and time of a documentation project. For this practical reason also, a systematic sampling procedure through the identification of types and inclusion of sufficient examples is most useful. For instance, we usually record the entirety of a festival that we attend on audio. However, it is not necessary to record the entire festival on video, since the dancing patterns (and other visual characteristics) actually vary very little throughout the performances of songs of the same type, e.g. during the performance of songs sung in the daytime. Thus a few good examples of each song type recorded on video (in addition to the audio recording) are sufficient to representatively document such a festival.

Things are quite different when attempting a representative documentation of informal events, i.e. events that are more spontaneous and less planned in the sense of the spontaneity parameter. We take as an example the event type scolding and look at how it can be defined and differentiated from other types. Unlike festival songs (which are always performed at a very precise place), scolding can take place at almost any **setting** or **scene**, and who the **participants** are may also vary quite freely. The **ends** seems to be what most accurately defines this event type, namely the purpose of informing someone of his wrong doing (and maybe re-establishing hierarchical social structure). The **act sequence** (message form and content) is also quite open to variation, although maybe certain lexical items and grammatical structures are more common in this type than in others. The **key** of a scolding event may be called serious (as opposed to mock). Its **instrumentalities** are spoken everyday

language. **Norms of interaction** include that among (sober) adults scolding is only appropriate in rare cases of serious anger.

Firstly, it is clear that an event type such as scolding is much harder to define and differentiate from other types than an event type such as a festival song. While there is never doubt about the type that a particular song belongs to, an informal event such as scolding may at the same time be insulting or advising, and there is no clear-cut limit between these types (although elicitation of translational equivalents of words such as 'scolding', 'insulting', etc., and related words in the language can help to define these).[8] This is clearly related to the spontaneity parameter: the well-plannedness of an event type such as a festival song results in homogeneous performances of each instance of this type according to ethnographic and linguistic characteristics such as those mentioned above, while the unplanned nature of an event such as scolding results in the possible variation of many such characteristics.

A second, maybe more important point, is that informal events are in general less amenable for recording in natural circumstances. This is clearly the case with scolding, i.e. real instances where a speaker seriously scolds another. In probably most speech communities, it would be inappropriate to record such an event and archive or even publish this recording in the context of a language documentation without infringing the privacy rights of the participants. Speakers can, however, be asked to act out such an event.[9] The shared experience of our and other documentation projects has been that speakers are very good at this, i.e. the acted out versions of the event closely resemble what could have been observed but not recorded on another occasion. Thus, a representative sample of communicative event types can be enhanced by including acted out versions of events that are not possible to record in more natural circumstances, but — crucially — such a type needs to be identified previously through some systematics of event types.

The main conclusion from the discussion of these examples it that the possibilities of systematic inclusion of sufficient examples of communicative event types is more feasible for some event types than for others. Firstly, unplanned informal events such as scolding are less easily defined as recurrent types, and secondly, it is often not possible to document natural occurrences of them for ethical reasons. On the other hand, publicly performed, well planned events such as song types can be more easily described as a system of well-defined, recurrent types, and thus sufficient examples of each type can be recorded relatively easily.

[8] Also, it is not clear whether scolding should count as an event type on its own or whether it is something that may occur during another type of event

[9] Thanks to Patrick McConvell who suggested this technique to us.

6. Conclusion

The central aim of this paper was to explore the possibilities and limitations of developing and applying criteria that ensure representativeness of language documentation. Such criteria are based on an identification and classification of communicative event types and the inclusion of sufficient examples of each of these types. It was shown that such a classification is far from being self-apparent in the data, but rather an abstraction as the result of a detailed analysis, requiring deep knowledge of the language and culture as well as theoretical assumptions and justifications. While the problem of theory-dependence of documentation has been discussed for various aspects of language documentations, the theoretical grounding of the process of selecting events to be included in a language documentation within language documentations is an issue that deserves further consideration.

It was shown that approaches such as the ethnography of communication provide very useful parameters for developing criteria for systematic sampling, but that these parameters must be adapted in various ways for the purpose of ensuring a representative language documentation. A major challenge is the development of a theoretically grounded identification of minor vs. major event types, where the latter are more important to document to ensure representativeness. It was also shown that the possibilities of systematically sampling events are limited in the sense that it is more feasible for some events that for others.

In practice the usefulness of such procedures is, of course, not so much to avoid the inclusion of a particular recording (i.e. to avoid overrepresentation of an event type), since, as noted above, almost any recording is valuable if the alternative is no recording at all. The usefulness is rather to avoid underrepresentation of certain event types that could have been missed if no systematic procedure was applied, and to actively seek occasions to record them, or — if they can not be recorded in a more natural setting — ask speakers to act them out. Needless to repeat, applying a systematic procedure does in no way exclude also collecting data by other methods, such as convenience and externally motivated sampling.

While the advantages of systematic sampling procedures for a representative language documentation are clear, their possibilities are also limited in a number of ways, for instance by ethical issues. I want to stress here that the question of what is representative of a language or speech community is not just a methodological, but also an ideological and highly controversial issue. For instance, Foley (2003) rejects the inclusion of texts that exemplify normative grammar in a language documentation because they are seen as importing Western literate ideologies. On the other hand, speech communities may have developed a preference for edited texts, which are cleared of speech errors, repetitions, code-switching, etc., and thus arguably

exemplify normative grammar, and their inclusion may be preferred (Mosel 2006a). And, to end this paper with a more extreme examples: who decides and on what grounds whether to include drunk speech or conversation during violent fighting, which are unfortunately frequent event types in many speech communities, and should — on methodological grounds — be well represented in a language documentation?

References

Aschmann, Richard P. 1993. *Proto Witotoan*. Dallas: The Summer Institute of Linguistics and the University of Texas at Arlington.

Dixon, R. M. W. 1980. *The languages of Australia*. Cambridge: Cambridge University Press.

Foley, William A. 2003. Genre, register and language documentation in literate and preliterate communities. In *Language Documentation and Description, Volume 1*. 85-98.

Gilmore, Perry. 1985. Silence and sulking: Emotional displays in the classroom. In D. Tannen and M. Saville-Troike (eds.) *Perspectives on silence*, 139-162 Norwood, NJ: Ablex.

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

Himmelmann, Nikolaus P. 2006. Language documentation: what is it and what is it good for? In Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel (eds.) *Essentials of language documentation*, 1-30 Berlin: Mouton de Gruyter.

Hymes, Dell. 1971. *Foundations in sociolinguistics. The ethnography of communication*. Philadelphia: The University of Pennsylvania Press.

Jakobson, Roman. 1960. Closing statement: Linguistics and poetics. In *Style in language*, edited by T. Sebeok, 350-377. Cambridge: MIT Press, New York and London: J. Wiley and Sons.

Kaufman, Terence. 1994. Review of Proto Witotoan by Richard P. Aschmann, Arlington, TX: Summer Institute of Linguistics and University of Texas at Arlington, 1993. *Language* 70 (2): 379.

Lehmann, Christian. 2001. Language documentation. A program. In *Aspects of typology and universals*, edited by Walter Bisang. 84-97. Berlin: Akademie Verlag.

Lieb, Hans-Heinrich, and Sebastian Drude. 2000. Advanced glossing: A language documentation format. DoBeS Working Paper.

Mosel, Ulrike. 2006a. Fieldwork and community language work. In Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel (eds.) *Essentials of language documentation*, 67-85. Berlin: Mouton de Gruyter.

Mosel, Ulrike. 2006b. Sketch grammar. In Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel (eds.) *Essentials of language documentation*, 301-309. Berlin: Mouton de Gruyter.

Ochs, Elinor. 1979. Planned and unplanned discourse. In Talmy Givón (ed.) *Discourse and syntax*, 51-80. San Diego: Academic Press.

Saville-Troike, Muriel. 2003. *The ethnography of communication: an introduction*. Malden, MA: Blackwell.

Schultze-Berndt, Eva. 2006. Linguistic annotation. In Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel (eds.) *Essentials of language documentation*, 213-251. Berlin: Mouton de Gruyter.

Seifart, Frank 2000. *Grundfragen bei der Dokumentation bedrohter Sprachen*. Köln: Institut für Sprachwissenschaft der Universität zu Köln.

Seifart, Frank. In press. The prehistory of nominal classification in Witotoan languages. *International Journal of American Linguistics*.

Sherzer, Joel. 1983. *Kuna ways of speaking. An ethnographic perspective*. Austin: University of Texas Press.

Widlok, Thomas. 2004. Implications of ethnographic techniques for anthropological and linguistic theory. Paper given at the conference 'A World of Many Voices', Frankfurt/Main, September 2004.