

Language documentation as a challenge to description

Maarten Mous, Leiden University, m.mous@let.leidenuniv.nl

Introduction

Only a relatively small part of the world's languages have been analysed and thus our insight into the phenomenon of language can only be fragmentary. Yet many languages are disappearing or adapting rapidly to another, dominant language and there is ample reason to put all our efforts into expanding our database of linguistic knowledge. In the last decades we have seen several initiatives to this end and in particular funds have been set up to promote research on endangered languages.

We are at a crucial moment for data-oriented linguistics and in a better position than it has been for a while due to the following developments: (1) the endangered languages debate has drawn the attention to linguistic diversity, (2) it has also opened a discussion on the use, nature, and representation of linguistic data, (3) theory-oriented linguistics has come away from the opinion that the richness of the English language encompasses the world's linguistic diversity, and (4) there is tendency in theory-oriented linguistics to value performance rather than competence. In order to make full advantage of this momentum we need an integrative approach, combining data and theory orientation making sure that the insights from language analysis enter the linguistic debate. We need more linguists to take part in the data collection.

This situation and the new initiatives for endangered languages have resulted in new views on data collection: Language documentation is now seen as different from language description; there is an emphasis on use of data by others than the researcher and long after the researcher has collected the data; there is an emphasis on the technical aspects of data collection and storage; and on the role and the rights of the speech community in all of this. Other trends are the view that data should not be collected through elicitation, and there is growing interest in the collection of comparable data collected in a test situation. Thus, the awareness of the imminent loss of linguistic diversity has altered data-oriented linguistics and specifically it has resulted in reflection on data collection. These developments have been argued for in a number of recent publications, Himmelmann (1998), Gippert et al. (2006) to mention a few. There are a number of issues that I want to raise for which I fear that the insights of the experience of traditional fieldwork based linguistic description are at risk. These are the separation of documentation and description, the status of elicitation as a tool in data collection, the pressure to use modern media, the regard of the linguistic academic community for linguistic description and the decrease in attention for medium-sized languages. This is not so much arguing against the authors mentioned above who, I suspect, might agree on the points I raise here, but more against the general discourse that I sense in circles concerned with organisation of research and among linguists who are new to fieldwork based linguistics.

Before I address these issues I briefly explain how I view the academic activity of language description. Language description is a craft. A linguist working in the domain of language description needs craftsmanship in addition to analytical skills. The craftsmanship to hear and produce sounds, the craftsmanship to gain access to the ear of the native speakers, the craftsmanship to listen and interpret meta-linguistic statements of native speakers, the craftsmanship to elicit such statements, the craftsmanship to imagine a different semantic world and the ability to translate. Obviously these are intellectual activities requiring great intelligence. The reasons I portray them as craftsmanship are the following: First, you can only learn them by experience. Learning to do fieldwork is like learning how to drive a car: you don't learn it from a book, you need to do it yourself starting with an instructor at your side. Secondly, some people have more talent than others and one gets better by experience. Thirdly, you develop your own style and one can discuss the value of different styles but all have their virtues.

1. Documentation presupposes description

One of the recent developments is the insight that documentation is more than description, see Himmelmann (1998). This has made a very important and valuable impact on reflection about data. Various funding bodies emphasize the need to do documentation rather than just description. My worry is that the view prevails that we do documentation first and grammar writing later. Reading the websites of some of the funds of endangered languages (see the Hans Rausing Endangered Languages instructions for projects on their website and the DoBes website), they seem to propagate such an approach. In my view separating data collection from analysis would be a step back in history. We have gone beyond the phase of survey, taxonomy and cataloguing in descriptive linguistics and we are in the next phase of in-depth studies. In-depth study is only possible by analysis in the field, by setting hypotheses and testing them immediately in order to come up with new ones. The quality of a descriptive study is largely determined by the degree in which data collection was driven by analysis. If we separate data collection from analysis we go back to poorer language descriptions. But also the quality of the documentation is determined by the level of analysis. Without a thorough understanding of the language, our rendering, translations, annotations and glossing of "texts" can only be poor. Texts here means filmed speech.

2. Elicitation is inevitable

Another trend in the *discours* around data is the advice that one should not trust elicited data and thus not use elicitation as a tool in data collection. I do not see how that is possible. In fact, to me a course in field methods is primarily a training in elicitation. Elicitation is guided conversation about language data. It is not a questionnaire to be filled out. Also, when we collect texts, to work on the text with speakers of the language is elicitation. Texts are indispensable in descriptive work on a language but do not replace the use of elicitation. Of course one should not be naive about the psychological complex situation that involves elicitation. This situation contains the following elements: The linguist has a linguistic problem and prepares a mini test for the speaker in the hope that this test

will provide clues to the answer of her problem. The mini test is often but not always the translation of a particular sentence from the common language into the researched language. Such a test can take all sorts of forms. Sometimes it is asking for a reaction of a sentence produced by the linguist in the target language. Sometimes it is asking the speaker to complete an utterance or to produce a sentence containing particular forms. The mini test is interpreted by the speaker and thus the linguist has to take into account that there is this filter of interpretation. The linguist has to contemplate on the fact that the speaker has created a context in her/his mind. The speaker reacts on the basis of this interpretation and that reaction is again interpreted by the linguist and analysed as to what it tells her about the problem at hand and what new problems and insights it reveals. In a few seconds the linguist has to process this all and formulate a new mini-test. Speed is crucial here and speed comes with experience.

I am very happy with the growing insight that our view on a language can be seriously distorted if we do not make use of texts of different genres. But in addition to the growing use of texts there is also a tendency to use sets of test material: video clips of pictures that are used to obtain comparable data. These are very welcome additions to our instruments of data collection but my worry is that there are people who think that these tests by themselves provide adequate data to analyse the topic of the test. In my experience these tests can give a very incomplete and even distorted picture of the phenomenon under study. The paradox is that by addressing the issue of the complex psychology of the mini-test in elicitation by explicitly setting up an experiment, the effect of the complex psychology, which is still present, is no longer taken into account. Tests often give a wrong picture of a language. Burenhult (forthcoming) argues that the video elicitation materials for placement and removal events possibly give a biased result due to the culturally unfamiliar nature of the events shown; common specific verbs in Jahai (Mon-Khmer, Malay Peninsula) for culturally salient events of putting and removal do not appear in the data resulting from the video stimulus material that is prepared for typological comparison.

3. New media and new data formats take time away from analysis

Central in documentation is the web publication of searchable glossed texts with linked audio and video material. Many of the published texts are edited for slips of the tongues, false starts and other features of language production that are considered mistakes by speakers. Such editing is vital when publishing a written text and in order for such a text to be valuable for the speech community as a written source, often additional editing is in place. Texts as written medium and texts linked to audio and video recording are very different and both have their value. Writing down texts that are a precise rendering of the recorded and video-taped speech, glossing it, translating it and commenting on it is a time consuming activity. If we have to produce a large body of such texts in a (web)publishable form in a searchable structured format, that will take up a lot of research time despite the various programs that can help us with semi-automatic glossing. As a consequence there is less time left for writing up the analysis. The paradox has arisen that researchers have to devote much of their time not on collecting more material of the disappearing

language but to preparing their recordings for archiving in order to have a successful outcome of their project while this could in fact be done at a later stage. Archiving all the relevant data has rightfully become a must but this too takes a lot of time and the grants do not offer longer periods than before. An other issue is that the need to have top quality audio and video recordings has numerous repercussions for the way in which we conduct our fieldwork. These range from the impossibility for the linguist to do the recording and monitor the speech event at the same time to the view of the community on the linguist as an observer rather than as somebody participating.

4. Writing a PhD on an endangered language?

Doing a dissertation on an endangered languages can be a risky enterprise. There is a risk that the state of endangerment is such that a full analysis of the dying language has become particularly difficult and the researcher or her supervisor may conclude that the data do not suffice for a PhD dissertation. It is more convenient to do a project on an endangered language which is no longer vibrant, as a post-doc rather than as a PhD project. Working on an endangered language means working in a situation of language shift and thus the researcher needs intimate knowledge of the language that the people shift to as well. In Africa this is usually another African language, often not fully analysed either. A description of the language to which the people shift is the ideal preparation for a next project on the language that is in transition. There are other concerns about writing a PhD on an endangered language. Communities speaking endangered languages are often not strong in self-confidence and that can bring extra responsibilities for the researcher. When the speech community is very small it may be difficult for the researcher to find educated speakers with whom she can have meta-linguistic discussion. When I did my PhD it was the norm to choose a language with many speakers and I would not have wanted to miss the conversations with some highly educated mother tongue speakers. There are situations of language endangerment where these concerns do not apply. I do not want to plea against working on an endangered language for a PhD; my aim is point out the dangers and challenges.

In some linguistic academic circles a linguistic description of a language is not considered fit for a PhD thesis or of lesser value than a theoretically oriented work. Such a view can only be held by people who have not experienced the highly intellectual activity of a first linguistic analysis in combination with data collection. No other empirical science allows such a low esteem of analysis based on new data. This view should be eradicated. A PhD that consists of a (first) description (=analysis) of a language poses extra challenges to the student. The main challenge is time. It is impossible to write a complete grammar based on fieldwork in the time that is usually allotted for a PhD, particularly because fieldwork often takes away extra time in the form of coping with a different and often difficult environment (people fall ill) and a different culture (adaptation time is needed both in the field and when back). These two consideration, the view of some other linguists and the time factor have led some of us to propagate to write a PhD on a particular (and interesting) aspect of a language rather than aiming at an analysis that is as com-

plete as time allows. I have a problem with this solution. Not only because it strengthens the wrong view that writing a grammar is a lesser kind of linguistic achievement, but even more so because of the fact that it is not fruitful and even dangerous to analyse part of the language without a complete picture of the rest of the language. This will inevitably lead to mistakes. In particular, no fruitful analysis of any morpho-syntactic feature can be done without a complete understanding of the phonology. The greatest revolution in descriptive linguistics during the last century is that it has become the norm that a description of a language can only be done on the basis of a solid analysis of its sound system. In more and more linguistic training programs students choose either syntax or phonology as if these are independent disciplines. They are not when we describe a language. Doing morphosyntax without a complete understanding of the phonology and incorporating that analysis in the approach of the morphosyntax is a step back in history. It is not impossible to complete a first grammar of a language in a period of four years. In several academic centres, for example Leiden University and La Trobe University such grammars are very common PhD theses and widely recognised as important contributions to linguistics. Writing a grammar of language of an undescribed language is an excellent training for a career in linguistics. The described language will always provide materials for detailed articles on topics of wider linguistic impact; moreover, for the rest of her career the linguist will have a database of new illustrations for phenomena in whatever linguistics class she teaches.

5. The middle-size languages are in danger of being forgotten

In the long run Africa has a fair chance to keep its linguistic diversity. Mother tongues keep entering the educational system, language shift is often to another African language while the international language remains a second/third language. The mobile phone revolution has linked the spoken word to modern technology and the written world is giving way to the oral one. All these tendencies give hope for many African languages to survive but these are most like the languages of wider communication and the languages that are dominant in sub-national regions. Many of these languages are understudied and it is difficult to find grants for language description of languages that are not in danger. Yet that is precisely what we need for the survival of linguistic diversity in Africa in the future.

Conclusion

In conclusion I repeat what I have argued for in six theses:

1. Documentation *before* description is wrong
2. Elicitation is necessary
3. New media and new data formats take time away from analysis
4. It can be dangerous to work on an endangered language for your PhD
5. A description of a language is a holistic enterprise and an excellent one for a PhD
6. Work on the larger African languages is needed, also to ensure linguistic diversity

References

Burenhult, Niclas. forthcoming. The linguistic encoding of placement and removal events in Jahai.

Gippert, Jost, Nikolaus Himmelmann, Ulrike Mosel (eds.) 2006. *Essentials of Language Documentation*. Berlin: Mouton de Gruyter.

Himmelmann, Nikolaus 1998. Documentary and Descriptive Linguistics. *Linguistics* 36:161-195.

Hyman, Larry. 2004. Why describe African languages? In *Proceedings of the 4th World Congress of African Linguistics New Brunswick 2003*, ed. by Akinbiyi Akinlabi, Oluseye Adesola. Cologne: Rüdiger Köppe.

<http://www.mpi.nl/DOBES>

<http://www.hrelp.org>