elpublishing

Language Documentation and Description

ISSN 1740-6234

This article appears in: *Language Documentation and Description, vol 4*. Editor: Peter K. Austin

The role of metadata for translation and pragmatics in language documentation

HENRIK BERGQVIST

Cite this article: Henrik Bergqvist (2007). The role of metadata for translation and pragmatics in language documentation. In Peter K. Austin (ed.) *Language Documentation and Description, vol 4*. London: SOAS. pp. 163-173

Link to this article: http://www.elpublishing.org/PID/055

This electronic version first published: July 2014

The role of metadata for translation and pragmatics in language documentation

Henrik Bergqvist

1. Introduction

It is stating the obvious to say that translation relies on contextual information as much as it does on mapping the basic meaning of a word in one language, onto a corresponding word in another. An important portion of the vocabulary in any language also depends directly on the context in order to be accurately understood since it makes reference to certain aspects of the speech situation itself.

The problems faced in translating under-documented and sometimes endangered languages make a discussion about *how* the context of the speech situation can be included in documentation practices both necessary and urgent. The necessity is among other things due to the fact that language documentation is an *ongoing process* that to some degree is independent of one specific researcher's control, and his/her personal knowledge about the language.

What is needed is a format for including large amounts of explicitly stated information about the speech situation that can be accessed together with the archived digital resources that make up the primary data of the documentation. The natural label for this sort of attachable information is of course *metadata,* since it in fact is 'data about data'.

In the following section, the deficiencies of descriptive linguistics and how it contrasts with documentation practices are summarised with regard to providing sufficient contextual information for translation. In section 2, the concept of metadata is discussed with suggestions regarding a wider definition of the concept that deal with the requirements of translating pragmatically anchored speech. In section 3, what constitutes *situational metadata* is discussed, followed in section 4 by examples of some of the points raised in the previous sections from the perspective of my own documentation of Lakandon Maya. Finally, in section 5, the main points presented in the paper are summarised.

Henrik Bergqvist 2007. The role of metadata for translation and pragmatics in language documentation. In Peter K. Austin (ed.) *Language Documentation and Description, Vol 4*, 163-173. London: SOAS.

2. Attitudes and priorities in descriptive versus documentary linguistics

Traditionally, the defining concepts linguistic field workers have used to guide them in their work of describing a language have been *data* and *structure*. Descriptive linguistic field work has had one aim to overshadow all other goals, which is to bring back enough data and enough native and personal judgements about that data, so that the grammatical structure of the language can be discovered. Most other aspects of working as a stranger in a foreign culture have been pushed to the sidelines in the pursuit of this end.

The traditional way of carrying out field work stems from the idea that a language is an independent object that can be studied like any other object, even though it be a complex one. All that is needed is someone who speaks the language and who can provide the necessary data for the study of the structure of the language.

Some severe limitations are set on the kinds of work that can be carried out on data that has been collected under these conditions. Pragmatic information, especially, is severely restricted by the attitudes and practices that descriptive field workers have carried with them into the field.

Descriptive practices have failed to supply enough 'data about the data' (i.e. metadata) for research by a secondary researcher to be possible on pragmatically anchored data. The (usually short) ethnographic description that sometimes is included as a background to a descriptive analysis is, however valuable, not enough for purposes of translating such data.

Language use can only be viewed from the extra-linguistic context that constitutes both the background and the imperative for any speech act. It contains both an understanding of the proper circumstances for saying something, as well as the meaning of certain utterances from the situation in which they occur.

With very limited information on the specifics of the culturally conditioned context, comes limited means to draw any conclusions as to how that, often foreign, context affects and determines the speech practices that constitute the (at best) archived data.

Documentary linguistics adopts a different approach in this regard. Several defining features of documentary linguistics require that sufficient situational data is attached to digitally encoded materials in order to ensure their usefulness and transparency to other researchers and community members alike.

3. The role of metadata in translation and pragmatic research

For data collected as part of the documentation of a language to accommodate as many different kinds of research and uses as possible, it must be accompanied by a description of the immediate context surrounding the recording of primary linguistic data. The metadata should supply information about the situation for the recording of speech, and preferably also the about the type of speech event that is performed within the situation.

As observed by Nathan and Austin (2004:179), metadata is currently understood as "information that is attached to a file or a document for cataloguing purposes". This is a definition of "thin metadata". It is a top-down, minimalistic practice that ignores the diversity of information contained in the multiple levels of structure that make up the data.

A bottom-up process of constructing metadata, on the other hand, would explore the kinds of metadata that linguists, community members, language teachers, and other researchers have use for. The collection of *thick metadata* should, according to the authors, take place in the community at the time of recording (or in conjunction with it) since researchers are more likely to have access to those who were recorded and/or participated at the time, as well as to other relevant elements that may not be remembered, or included, in a later account (ibid. 180).

Metadata as a concept and a part of documentation practices must be enriched and broadened to include more than one level of information, specifically with regard to "information structure within a document" (ibid. 179).

The authors identify a "metadata-gap" between the minimalist cataloguing schemas and the rich descriptions that result from collecting and analysing field data, and argue that what is needed to support language documentation is a metadata methodology that is *flexible and explicitly articulated*, to accompany the layers of data that result from the documentation. In sum, metadata must be defined as being much more than the cataloguing of digital resources.

The focus in the present paper is on *situational metadata* since the concern here is to discuss the need for practices of metadata annotation to improve translation and pragmatic research using documentation materials. Situational information about the *location, time,* and *participants*, all constitute what can be labelled *deictic metadata***.** Deixis is obviously, not only an important topic in pragmatics, but also highly relevant for translation purposes, and deictic metadata is a requirement for work on both.

4. Situational metadata

Minimally, situational metadata should include information about i) the speech participants, ii) the location of the recorded speech event, and iii) the time when it was recorded.

As for the speech participants, every person that in some way is relevant to the documentation process, should be identified and presented[1]. The kinship relations between the speech participants is of course a necessary part of such a description. The documenter (i.e. the linguist) should not be left out from this account. An attempt to identify the status of the linguist within the community along with a description of the stage of research that the documentation has reached at the time of recording is also relevant. Not as interesting extra-information, but as a genuine part of the documentation process that enriches the data and opens it up for more than traditional grammatical analysis.

Information about the place and time of a recording in a thick metadata description, would include not only the name of the location (e.g.. Enrique's kitchen), the time of day (noon), but also information about other activities that take place in the same location, thus revealing something about the attitudes of the speakers towards the chosen location for answering questions or recording speech. The description of the location should be given in a manner that relates it to other specific locations and areas that the speech participants (and the documenter) have access to, or are excluded from.

The time of the recording should equally be defined both with regard to the diurnal span and other relevant frames of reference, such as ritual calendars and the agricultural cycle. How a certain time of day relates to day-to-day activities and routines such as eating, working, playing, and resting may also be included.

The granularity of situational metadata must be as fine as the researcher can afford. There should not be a limit on the amount of metadata that the primary documenter supplies together with collected 'primary' data. The aim must be to let as little data as possible be left on a level of isolation that prevents a clear view of the context of the recorded speech event. The open ended-ness of documentation is surely the strongest argument against drawing the lines too narrowly in this respect.

The limits in time and money set on the initial stages of a documentation project usually mean that a lot of material is left hanging without being

[1] The identification and presentation of a participant of the documentation project can of course be done in a way that veils his/her true identity, if that is what the person wishes. What is important is to make all participants visible *as participants* and as being in some relation to each other as well as to the documentation itself.

properly understood or attended to. It is furthermore unfeasible to ask that all the specific elements connected to the recording of a speech event be properly understood in those early stages. The underlying attitudes towards the (often artificial) speech situation; the constraints that exist on the topics of speech; the impact of the presence of researchers on the material they are gathering, and a number of other things that can be connected to attitudes, beliefs, viewpoints, and concerns of the speakers, are often not immediately obvious to the researcher. By increasing situational, deictic information to go with archived recordings of speech, issues such as these may be better understood from later research efforts.

The confusion that field workers sense when, for instance, they encounter a foreign narrative tradition comes from not knowing the prerequisites of that tradition. After some questioning and pondering many of the issues that initially surprised, or confused the linguist may be resolved, but we have to assume that the foreign-ness of any tradition is present on several levels and that some parts of that tradition may never present itself to the primary researcher. That is, it may never come up as something unfamiliar that needs to be resolved or disambiguated but it may nevertheless be pertinent to the quality and interpretation of the data in ways that linguists fail to see because of the attitudes they bring with them into the field and because of the questions that they decide to ask.

The idea that "a language belongs to the person documenting it" must be abandoned to allow the possibility of a continued documentation process.

5. Some examples from my own work: deictic time in Lakandon

There are some important consequences for translation and meaning retrieval if situational metadata is broadened in the ways described above. An example of the need for information about the time when a story was recorded, is found in the paper by Evans and Sasse (this volume) where the use of the word for 'now' or 'nowadays, today' in a Dalabon narrative, initially confused the translator until a connection was made possible from a comment by one of the narrators relating the temperature of the season they were in at the time of recording, with the conditions of living like a crocodile under water. A phrase that did not make sense at first was made transparent by taking into account a characteristic property of the time period when the story was told (ibid. §3.2.5).

5.1 Reference to past events in Lakandon Maya: *7uhch* and *ka7ch(ik)/kuhch*

As part of my own documentation of Lakandon Maya, an under-documented Yukatekan language spoken in the lowlands of southeastern Chiapas, I am currently trying to analyse and understand the use and meaning of deictic time words. I am interested to know what the motivations and conditions are for their use since they show some interesting patterns of distribution and meaning that seem to defy an exclusively temporal interpretation (Bergqvist, forthcoming).

One of the conclusions that William Hanks (1990) draws from his investigation of deictics in Yukatek is that there is a "tendency towards proportionality across categories" (Hanks 1990: 487); i.e. the features that are relevant for the description of one deictic category are probably also present in another category. Given the common origin of Yukatek and Lakandon and their recent separation into separate languages (no more than 500 years), it is reasonable to assume some degree of preservation of the semantics and structural make-up of Yukatek deictic forms, in Lakandon.

From a structural point of view these assumptions are confirmed in that the forms attested for Yukatek are almost entirely present in Lakandon with regard to the deictics for *space, presentation* (ostensive deictics), *nominals*, and *person*. If the semantics of the investigated categories are equally conserved, is a question that remains to be answered in full.

However, several of the semantic parameters that are relevant for the description of the ostensive deictics forms in Yukatek, are present in the meaning of deictic forms for time in Lakandon. For example, there are two distinct forms that make reference to a 'past' event, that are differentiated by *knowledge asymmetries* between the speaker and the addressee, similar to what Hanks reports with regard to the speaker's and the addressee's symmetries of access to an object, in the forms *je7ra7* and *je7ro7*.

One form, *7uhch* ('before', 'long ago') indicates that the speaker believes the information he is presenting to be previously unknown to the addressee. The other form *kuhch* or *ka7chik* (depending on whether the speaker is speaking the southern or the northern dialect), refers to an event or state that the speaker has reason to assume, usually from direct evidence, is known to the addressee.

An illustration of the different meanings connected to the two forms can be viewed in a comparison between them. The first example is from a conversation between one of my main consultants, Enrique, and a visiting non-relative. Here Enrique tries to clarify to the visitor what he himself has

said on a previous occasion regarding a rather complicated matter of a broken water pipeline:[2]

(1) *ma7* *7inw-a7r-aj* *raj-i7* *[ka7] yäx juhntaj* *7uhch*
NEG 1SG.A-say-COM it-ANA [when]first meeting before_spkr
'I didn't say that at the last meeting.' (HB041023_1EChK_7)

By using *7uhch,* Enrique states his personal perspective on what he said at the meeting in question. In addition, the stance adopted by Enrique is in contradiction to what the visiting man had suggested at the beginning of the conversation.

A different perspective is present in an explanatory narrative where Enrique is explaining the Lakandones' interpretations of dreams (HB040905_2EChK_7). Present is also Enrique's son Enrique K'ay Yum, and when Enrique loses his line of thought and hesitates on what to tell me next, he asks his son to help him by restating what he has said only a little while ago:

(2) a b'ay t-aw-a7r-aj ka7ch-ik
what CP-2SG.A-say-COM before_adr-ADVR
'What did you say before?'

b *chäkäw*
hot
'Fever'

2 The phonemic orthography used in this presentation is identical to the one used by the PDLMA project (dirs. T. Kaufman, J. Justeson, R. Zavala), and closely approximates the orthography established by the Guatemalan Academy of Mayan Languages: /ä/ is a "schwa", /7/ is a glottal stop, /j/ is a voiceless glottal fricative, /h/ indicates high tone (two tone distinction; high and low, (low unmarked)). Abbreviations used in the glosses are 1: first person, 2: second person, 3: third person, .A: setA (ergative marker), .B: setB (absolutive marker), ADVR: adverbial marker, ANA: anaphorical reference marker, CAUS: causative, COM: completive status marker, CP: completive aspect, DIST: distal terminal deictic (TD) marker, DEP: dependent status marker, DET: determiner, INC: incompletive (plain status), IND: independent pronoun, LOC: locative, NEG: negative, NOM: nominal suffix, OST: ostensive initial deictic (ID) form, PL: plural, PN: personal name, PROX: proximal TD marker, REF: referential/anaphorical, REFL: reflexive, SG: singular, TN: toponym, TOP: topic marker, TR: transitiviser.

c *7a-ma7* *chäkäw* *b'a7ikin* *t-aw-a7r-aj* *ka7ch-ik*
DET-NEG Hot which.one CP-2SG.A-say-COM before_adr-ADVR
'Fever wasn't what you said'

d *k'uxu7*
achiote
'Achiote'

e *7a-k'uxu7* *la-je7 [...]*
DET-achiote that-OST
'Achiote, that's it (what you said)'

In (2a) Enrique requests a repetition of information that the addressee has already uttered. However, the perspective of the speaker is prevalent since Enrique disqualifies the response he gets by disagreeing with his son and asking for another utterance to be repeated. By keeping *ka7chik*, Enrique maintains a perspective where information is known to the addressee and that he has already told the speaker.

In descriptions of the forms (*ka7ch*) in Yukatek (Bohnemeyer 1998) and (*kuch*) in a third Yukatekan language, Itzaj (Hofling 2000), the interpretation is limited to purely temporal parameters of *anteriority*. Bohnemeyer calls *ka7ch* ('previously') a "topic-time shifter" that can be explained simply by being contrasted to *b'ejohra* ('now'). Hofling labels *kuch* as "counter-factual" indicating an event or state that is contrary to what is within the present time.

From the examples that both authors provide, it is impossible to know if the asymmetry parameter is present, simply because the examples they cite are left without any context or reference to who uttered them and to whom they were being directed.

Perhaps because of disregarding pragmatic information, Bohnemeyer states that *ka7ch* sometimes can be substituted, or co-occur, with *7uhch* (also present in Yukatek), but that the two forms seem to contain no difference in meaning (Bohnemeyer 1998: 311).

5.2 Reference between events using *b'eje7* 'now'

Another temporal deictic form that contains semantic features of knowledge assymetry is *b'aje7(re7)* ('now'). The use of *b'aje7(re7)* is motivated by the same asymmetry of knowledge that what we saw in the 'past' form, *7uhch,* where the speaker presents information that he has reason to assume is unknown to the addressee. It also appears that *b'aje7(re7)* makes 'ostensive' reference to events and states in a way that is completely parallel to what is reported for the ostensive forms in Yukatek with regard to objects or persons, a function that seems to be present in the same ostensive forms in Lakandon.

In a story by a female speaker, CChNK, about the first time she gave birth, *b'aje7* is used to point out a contrasting event to the main events of her story. The recording was done by Una Canger (1970), in San Christóbal de las Casas, Chiapas, as part of a project to make root-dictionaries for Mayan languages. Canger had not transcribed, nor translated the story when she gave the digitised recording to me. This was done by me with the help of a native speaker in the field.

The contrast that *b'aje7* points out is between the first time the speaker gave birth and another (second?) time when the experience of giving birth was less distressing and painful:

(3)	*b'aje7*	*7a-je7*	*juntuhr*	*uhch-o7*	*7a-ray*	*ma7*
	now	DET-this	other	before_spkr-DIST	DET-it	NEG

'Now, that other one, not this one'

What initially confused me in translating the passage was that the speaker emphasised so much being in pain, but that I interpreted the phrase in (3) to mean that she was not in pain now (at the time of telling the story) but only at the time of giving birth to her child.

However, to support my most recent understanding of the utterance, where CChNK is switching to talk about another comparable event, it is important to know if the speaker has given birth to more than one child. CChNK has five children but one must also know how old they all are in order to know who had been born at the time of recording the story.

The phrase in (3) is very hard to understand and translate without information of this kind. A more correct translation of (3) would read: 'This time, this other (occasion), not that one (that I was talking about)'.

6. Summary and conclusions

The goals and practices of descriptive linguistics are insufficient if collected language materials are to be accessible and indeed useful to other researchers than the primary documenter. An important aspect of language documentation that separates it from descriptive practices is the way metadata is shaped and attached.

This is especially so in the case of situational metadata which, regardless of which media it accompanies, is essential to rule out 'guesswork' as a method to translate endangered languages. The translation of under-documented languages must be given more time and effort than the initial stages of documentation usually can afford. This can only be achieved by making the information supporting recordings of speech explicit to multiple parties.

There may eventually be little use for some of the extra-linguistic information included in a thick metadata description with regard to how it affects translation and linguistic analysis. But since another difference between descriptive- and documentary linguistics is the *separation of data collection from data analyses*, this eventuality can not be used as an argument for leaving the collected data without information that potentially will make it a great deal richer and more open to future research.

The accessibility of documentation materials (through the process of archiving) to researchers and community members alike, and the ongoing-ness of a documentation project, also makes it reasonable to request a certain amount of redundancy of information for a realisation of the goals of documentary linguistics.

The suggestions presented here are not a roundabout way of asking documenters to be anthropologists as well. Nor is my suggestion that the linguist should include a description of 'the entire world'. However, for a documentation corpus to be useful for pragmatic research and open to a ongoing process of translation, bare language and cataloguing data is not enough.

In the continuing definition of documentary linguistics, issues such as these must be included both as an argument supporting documentary linguistics as a separate branch and as a justification for the changes in attitude that are needed to make the goals of documentary linguistics, as defined by Himmelmann (1998) inter alia, possible.

Documentary linguistics should not be defined by the practical limitations connected to carrying out field work. It must be shaped by the goals and aspirations that the participants in language documentation set for themselves.

References

Bergqvist, Henrik (forthcoming). Semantics of temporal deictics in Lakandon Maya. In Proceedings from CILLA II (Congreso de Idiomas Indígenas de Latinoamérica). University of Texas at Austin 27th-29th October 2005.

Bohnemeyer, Jürgen 1998. *Time Relations in Discourse: Evidence from a comparative Approach to Yucatec Maya.* PhD Dissertation, Katholieke Universiteit.

Canger, Una 1970. Lacandón recordings. Digitized at the University of Copenhagen.

Evans, Nick and Hans-Jürgen Sasse 2004. Searching for meaning in the Library of Babel: field semantics and problems of digital archiving. This volume.

Hanks, William F. 1990. *Referential practice: language and lived space among the Maya.* Chicago: University of Chicago Press.

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

Hofling, Charles A. 2000. *Itzaj Maya grammar (with Félix Fernando Tesucún)*. Salt Lake City: University of Utah Press.

Nathan, David and Peter Austin 2004. Reconceiving metadata: language documentation through thick and thin. In Peter K. Austin (ed.) *Language Documentation and Description, Vol 2*, 179-187. London: School of Oriental and African Studies.