11

Language Documentation & Conservation Special Publication No. 3 (August 2012):
Potentials of Language Documentation: Methods, Analyses, and Utilization,
ed. by Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann,
Dagmar Jung, Anna Margetts, and Paul Trilsbeek, pp. 83–89
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4520

How to measure frequency? Different ways of counting ergatives in Chintang (Tibeto-Burman, Nepal) and their implications

Sabine Stoll and Balthasar Bickel
University of Zürich

The frequency of linguistic phenomena is standardly measured relative to some structurally defined unit (e.g. per 1,000 words or per clause). Drawing on a case study on the acquisition of ergativity by children in Chintang, an endangered Tibeto-Burman language of Nepal, we propose that from a psycholinguistic point of view, it is sometimes necessary to measure frequencies relative to the length of the time windows within which speakers and hearers use the language, rather than relative to structurally defined units. This approach requires that corpus design control for recording length and that transcripts be systematically linked to timestamps in the audiovisual signal.

1. INTRODUCTION. Both in historical linguistics and language acquisition research, frequency is generally assumed to be one of the most important features influencing language development (e.g. Bybee & Hopper 2001). One of the main assumptions of the usage-based approach is that distributions of patterns, i.e. frequency distributions and repetitions, play a key role in language change and language learning, underlying the gradual emergence of constructions diachronically (e.g. Hopper 1988) and developmentally (e.g. Tomasello 2003).

However, since frequency is a relational measure, any counting is meaningless unless we have a unit over which we can reasonably assume that the relevant items are tracked by speakers and hearers when processing language: we can count phenomena per linguistic unit (words, phrases, clauses etc.), per non-linguistic context and genre, per content unit (the choice of specific topics), or per time unit (in, say, minutes of speech or hours of conversation). It is unclear *a priori* what kind of unit is most useful for a given research question. Although the choice of counting unit has fundamental consequences on the results, this issue has received surprisingly little attention. The issue is particularly pressing, however, when we design and compile relatively small corpora, such as corpora of spoken and endangered languages, because the choice of counting unit predetermines the kinds of factors one needs to consider: to what extent is it important to balance or control for content types, recording time length, number of words, etc., and which of these is important for what research purpose?

In this paper we discuss some of the consequences of choosing among a variety of counting units. We exemplify these issues with a study on the role of frequency in the acquisition of ergative case in Chintang (`ISO639.3:ctn`, Tibeto-Burman/Sino-Tibetan, Eastern Nepal), based on a corpus that we compiled as part of a DoBeS project.[1] A key advantage of the corpus is that it is systematically linked to time stamps in the audiovisual recordings, and this makes it possible to consider not only counting units that are defined in terms of grammar or content but also in terms of time flow.

2. DATA. Chintang is a polysynthetic language spoken in a village in Eastern Nepal by about 6,000 people, who are all bilingual in Nepali, the *lingua franca* of Nepal (e.g. Bickel et al. 2007, 2010, Stoll et al. 2012). The language is endangered, but there is still a substantial number of children who learn the language as their first language. Our study is based on a longitudinal language acquisition corpus of 4 children learning Chintang (all from different families). Two children were aged 2 years and two children aged 3 years at the beginning of the study. The children were recorded over a period of 18 months for about 4 hours per month, while playing in their natural environment (mostly outdoors), with many different interlocutors around, both children and adults. A minimum of one and a half hours of recordings per month were used for the present study. The data were transcribed, translated, morphologically glossed, and tagged for part of speech properties (for more information see http://www.spw.uzh.ch/clrp). Figure 1 shows the amount of data available for the different children and recording sessions.

3. ERGATIVE MARKING IN CHINTANG. Ergative case in Chintang is distributed along a split system conditioned by person. The ergative marker (*-ŋa*) occurs obligatorily only with third person noun phrases. For first and second person the marker is optional, and for first person exclusive it is ungrammatical. Additional complications come from the fact that arguments are very frequently dropped in Chintang discourse (Stoll et al. 2012) and that the same case form *-ŋa* also doubles as an instrumental and an ablative marker (Bickel et al. 2010). As a result, ergative case does not seem to have a very high cue validity in Bates & MacWhinney's sense, even in third person contexts. This would make the acquisition of ergative case particularly challenging and difficult to account for.

But the question arises whether this impression of low cue validity is in fact empirically justified. In order to examine this, we need to chart the actual distributions in the speech of native adult speakers. In the following we analyze the adult speech surrounding our target children in the corpus.

4. MEASURING FREQUENCY. As noted in the introduction, the key issue in exploring frequencies is the choice of unit over which we count frequency. Usually there is more than one option. Each option leads to very different results, but more importantly, each option

[1] The data are available in the DoBeS archive, http://corpus1.mpi.nl. We use a snapshot of the corpus from October 2010, with a total size of ca. 280,000 words. Development of the corpus was made possible by a DoBeS grant (PI Balthasar Bickel) and a Dilthey fellowship to Sabine Stoll, both from the Volkswagen Foundation. Our research is embedded in the Chintang Language Research Program (http://www.spw.uzh.ch/clrp), and we are grateful to our colleagues in the program, especially Sebastian Sauppe, Taras Zakharko, and Robert Schikowski for help in preparation of the corpus for the present study. All corpus analyses were performed in R (R Development Core Team 2011) and visualized using the package `lattice` (Sarkar 2008).

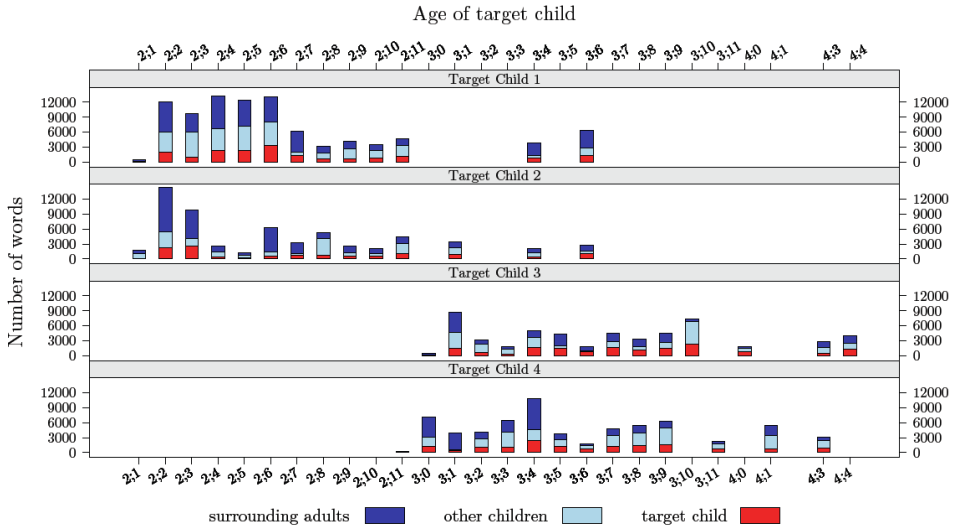

FIGURE 1: Distribution of data in number of words.

also makes strong but implicit assumptions about language processing and memory, both when learning a language for the first time in acquisition and when replacing one variant with another in language change. While this is not the place to review the psychological literature on these assumptions, we present a case study in the following that explores the general kinds of assumptions and overall results that are tied to four specific ways of measuring frequencies. We take as an example the acquisition of ergative case in Chintang.

4.1. RAW NUMBERS PER AGE IN MONTHS. A first relational option is the use of raw numbers per age, e.g. per month of age. This measure is rarely chosen because it is probably not very useful in most contexts without knowing what these numbers relate to. It obviously makes a huge difference if we find 5 instances in a corpus of 1,000 words or in a corpus of 10,000 words. Thus, the relational component is crucial for evaluating the numbers, and it should be explicitly stated. This is so in the options for counting that we consider in the following.

4.2. ERGATIVES PER WORD. Another option counts how often per word unit the ergative would occur, i.e. the proportion of words with an ergative marking. This would give us an impression of how often a child hears such a marker independently of its syntax or semantics. Results are shown in Figure 2.

If we used this measure, ergative marking would indeed appear to be exceedingly rare, never exceeding about .03%. However, to the extent that we would not want to assume that children parse language completely without any semantic or structural analysis, this measure might not be very revealing. Further, counting simply ergatives per word ignores the fact that ergatives can only occur in certain syntactic contexts: they are limited to noun phrases functioning as transitive agent ('A') arguments of transitive verbs. This brings up another relational type of counting ergatives.

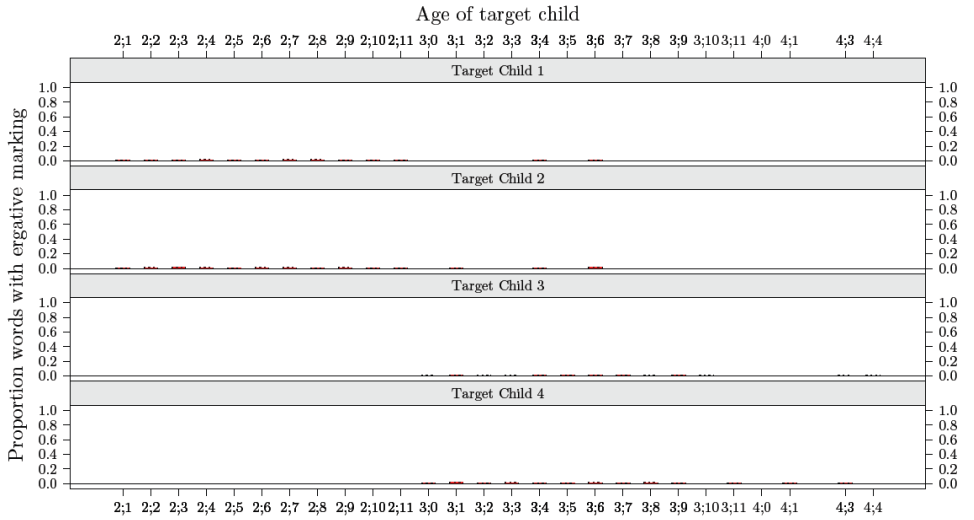

FIGURE 2: Proportion of words with ergative marking in adult speech surrounding the target children in the study

4.3. ERGATIVES PER TRANSITIVE VERB. Figure 3 illustrates the proportions of transitive verbs with an ergative marker per child and age. We exclude from this the occurrence of the same marker in an instrumental or ablative function, i.e. we limit our attention to the transitive A role.[2] On this count too, ergatives seem to be rare, although with a maximum of 11%, not as rare as when counting ergatives per word (.03%, Figure 2). At any rate, this would still be in line with the expectations derived from purely structural considerations.

However, counting ergatives per transitive verbs begs a number of questions: why should we choose all transitive verbs, rather than only verbs that are actually used with transitive syntax (cf. Note 2)? If we choose all transitive verbs, should we include A arguments across all persons, or should we limit our attention to third persons since it is only here that ergatives are compulsory? Regardless of what answer we give, it will invariably make the psychologically very strong assumption that the child has abstract knowledge over all these features of grammar (such as lexical vs. syntactic transitivity, or person categories), i.e. that the child parses the input on the basis of a fairly fine-grained distributional analysis. It is not at all clear, however, whether such an assumption is indeed warranted. Similar issues arise when considering the psychological bases on which speakers, regardless of their age, engage in language change: when new forms are innovated and especially when (as is often the case) forms are extended to new contexts, it is unclear whether and to what extent speakers make consistent distributional assumptions about the context from which the innovation starts.

[2] Transitive verbs can also be used in detransitivized constructions, where the A argument receives nominative case. For present purposes we gloss over these different uses and only consider the bare opportunity for ergative case marking which is associated with every transitive verb. For further discussion see Bickel et al. (2010), Schikowski et al. (2010).

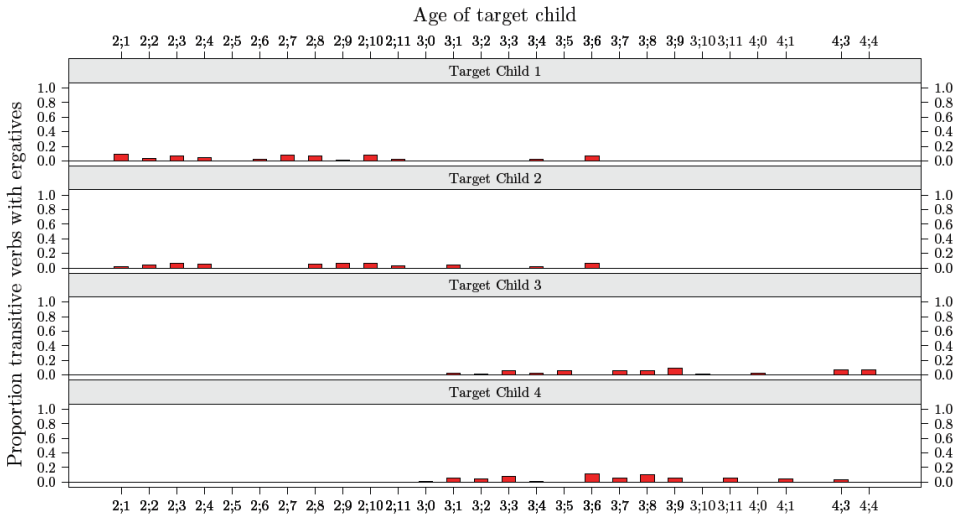

FIGURE 3: Proportion of transitive verbs combining with an A argument marked by ergative case in adult speech surrounding the target children in the study

An additional difficulty in Chintang concerns the fact that agents are often named in isolation, with the verb dropped (e.g. because it was mentioned in a previous conversational turn). These cases are excluded when counting ergatives per verb, but ergatives in isolation might provide key contexts that help children learn their use.

4.4. ERGATIVES PER TIME UNIT. Under this approach we consider the density in which ergatives are offered to children (or hearers more generally). Density of occurrence is arguably a psychologically important unit since it directly relates to well-known memory demands on processing and learning. Figure 4 shows the counts of ergatives per hour of speech. This includes all ergatives, regardless of their context.

In stark contrast to all previous frequency counts, counting ergatives per hour, i.e. in terms of the density of occurrence, suggests that the number of cases that a child hears is not so small after all. Children hear the ergative on average every two minutes (30 occurrences per hour), sometimes even every minute. To the extent that density of occurrence is psychologically relevant, this relatively high density would seem to facilitate the learning process considerably.

5. CONCLUSIONS. The present study suggests a distinction between two types of frequency measures. One measure relies on the frequency of X relative to the structural opportunity for X. This is the standard in corpus linguistics and also in usage-based theory. However, the psychological relevance of this type of frequency measure is unclear because it relies on very strong assumptions about the extent to which the 'opportunity for X' is in fact known and taken into account by hearers when learning a language or when being involved in language change.

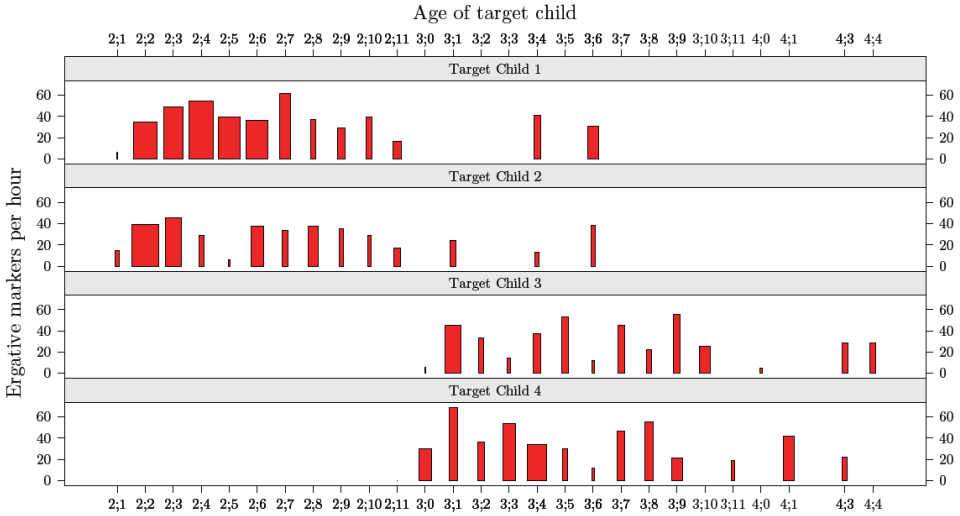

FIGURE 4: Proportion of ergatives per hour in adult speech surrounding the target children in the study. (Bar width is proportional to corpus size in number of words.)

An alternative measure relies on the frequency of X within a given time window and aims at estimating the density of occurrence of X. This measure directly relates to the demands on memory and processing that are relevant for language learners. This measure makes minimal assumptions about the level of analysis that a hearer uses, and at the same time, it gives an impression of how often a hearer is confronted with the feature in question.

For such a measure to be applicable, corpora need to control not only for genres, register, contexts, etc. (as emphasized by Lüdeling, this volume), but also for recording length. For this to be possible, transcripts need to be systematically linked to timestamps in the audiovisual signal.

REFERENCES

Bates, Elizabeth & Brian MacWhinney. 1982. Functionalist approaches to grammar. In Eric Wanner & Lila R. Gleitman (eds.), *Language acquisition: the state of the art*, 173–218. Cambridge: Cambridge University Press.

Bickel, Balthasar, Goma Banjade, Martin Gaenszle, Elena Lieven, Netra Paudyal, Ichchha P. Rai, Manoj Rai, Novel K. Rai & Sabine Stoll. 2007. Free prefix ordering in Chintang. *Language* 83. 43 – 73.

Bickel, Balthasar, Manoj Rai, Netra Paudyal, Goma Banjade, Toya Nath Bhatta, Martin Gaenszle, Elena Lieven, Iccha Purna Rai, Novel K. Rai & Sabine Stoll. 2010. The syntax of three-argument verbs in Chintang and Belhare (Southeastern Kiranti). In Andrej Malchukov, Martin Haspelmath & Bernard Comrie (eds.), *Studies in ditransitive constructions: a comparative handbook*, 382–408. Berlin: Mouton de Gruyter.

Bybee, Joan L. & Paul J. Hopper. 2001. Introduction. In Joan L. Bybee & Paul J. Hopper (eds.), *Frequency and the emergence of linguistic structure*, 1–24. Amsterdam: Benjamins.

Hopper, Paul. 1988. Emergent grammar and the a priori grammar postulate. In Deborah Tannen (ed.), *Linguistics in context*, 117–134. Norwood, NJ: Ablex.

Lüdeling, Anke. this volume. A corpus linguistics perspective on language documentation, data, and the challenge of small corpora.

R Development Core Team. 2011. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://r-project.org.

Sarkar, Depayan. 2008. *Lattice: Multivariate data visualization with R*. Berlin: Springer.

Schikowski, Robert, Netra P. Paudyal & Balthasar Bickel. 2010. Fluid transitivity in Chintang. Paper presented at the workshop on Valency Classes, MPI for Evolutionary Anthropology, Leipzig, 21 August 2010 [http://www.spw.uzh.ch/schikowski/work/2010-fluid-transitivity.pdf] (21 March, 2012).

Stoll, Sabine, Balthasar Bickel, Elena Lieven, Goma Banjade, Toya Nath Bhatta, Martin Gaenszle, Netra P. Paudyal, Judith Pettigrew, Ichchha P. Rai, Manoj Rai & Novel Kishore Rai. 2012. Nouns and verbs in Chintang: children's usage and surrounding adult speech. *Journal of Child Language* 39. 284 – 321.

Tomasello, Michael. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, Mass.: Harvard University Press.

Sabine Stoll
sabine.stoll@uzh.ch

Balthasar Bickel
balthasar.bickel@uzh.ch