From text to typology: towards implementing quantitative typology on corpora from endangered languages

Geoffrey Haig & Stefan Schnell

Proceedings of Conference on

Language Documentation & Linguistic Theory 2

Edited by Peter K. Austin, Oliver Bond, Monik Charette, David Nathan & Peter Sells

13-14 November 2009 School of Oriental and African Studies, University of London

Hans Rausing Endangered Languages Project
Department of Linguistics
School of Oriental and African Studies
Thornhaugh Street, Russell Square
London WC1H 0XG
United Kingdom

Department of Linguistics:
Tel: +44-20-7898-4640
Fax: +44-20-7898-4679
linguistics@soas.ac.uk
http://www.soas.ac.uk/academics/departments/linguistics

Hans Rausing Endangered Languages Project:
Tel: +44-20-7898-4578
Fax. +44-20-7898-4349
elap@soas.ac.uk
http://www.hrelp.org

© 2009 Geoffrey Haig & Stefan Schnell

No part of this publication may be reproduced, stored in a retrieval system, or transmitted, on any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the author(s) of that part of the publication, except as permitted by UK copyright law.

ISBN: 978-0-7286-0392-9

This publication can be cited as:

Geoffrey Haig & Stefan Schnell. 2009. From text to typology: towards implementing quantitative typology on corpora from endangered languages. In Peter K. Austin, Oliver Bond, Monik Charette, David Nathan & Peter Sells (eds) *Proceedings of Conference on Language Documentation and Linguistic Theory 2*. London: SOAS.

or:

Geoffrey Haig & Stefan Schnell. 2009. From text to typology: towards implementing quantitative typology on corpora from endangered languages. In Peter K. Austin, Oliver Bond, Monik Charette, David Nathan & Peter Sells (eds) *Proceedings of Conference on Language Documentation and Linguistic Theory 2*. London: SOAS. www.hrelp.org/eprints/ldlt2_12.pdf

From text to typology: towards implementing quantitative typology on corpora from endangered languages

GEOFFREY HAIG & STEFAN SCHNELL
University of Kiel

1. INTRODUCTION

Linguistic typology still predominantly relies on data from grammars, that is, data that are pre-analysed and to some extent selective. These data are often interpreted as evidence for the mere presence or absence of abstract structures in a given language, which in turn yield traditional typological generalizations: language A 'has' prenominal relative clauses, language B 'has' differential object marking, and so on. But it is becoming increasingly evident that there is also a cross-linguistic bedrock of commonalities at the level of statistically significant distributions of grammatical features across discourse. These patterns are generally ignored, or obscured, in descriptive grammars and have only emerged through the comparative investigation of natural discourse. Pioneering work in this area has been undertaken by John Du Bois (e.g. Du Bois 1987, Du Bois et al. 2003) and Balthasar Bickel and his associates (e.g. Bickel 2003, Stoll & Bickel 2009) which we discuss in the next section.

The current global initiatives for language documentation (e.g. DoBeS, HRELP, ELF) have (among many other things), produced an unprecedented number of corpora of transcribed spoken language, often from small and typologically unusual language communities. They provide a fertile and as yet largely untapped resource for text-based, as opposed to grammar-based typology. In this paper, we introduce a one-tier model for morpho-syntactic annotation, intended for application to a wide range of typologically diverse languages. The annotation captures information regarding (i) animacy distinctions, (ii) coding distinctions, and (iii) syntactic functions of major constituents, using a relatively simple and flexible system. The resultant annotations will form the basis for quantitative comparative analysis, and can be used to address such questions as how animacy distinctions affect coding properties (zero vs. bound vs. pronominal etc.), which syntactic functions are most prone to which kind of coding (cf. Du Bois' PREFERRED ARGUMENT STRUCTURE (PAS) and related 'soft' constraints), and to differences in the extent that languages overtly realize verbal arguments (Bickel's REFERENTIAL DENSITY (RD)). The system aims at reaching a compromise between the basic annotation system of Du Bois (1987), and the more complex tagging systems currently used in tagging corpora of standardized languages (e.g. the tagging system employed by the British National Corpus). The system has been trialed on corpora from two languages, Gorani (West Iranian) and Vera'a (Austronesian, Oceanic), for which currently approx. 400 clause units have been annotated, and is intended to be extended to further languages from other documentation projects.

2. PREFERRED ARGUMENT STRUCTURE AND REFERENTIAL DENSITY

Typological research on statistical regularities in grammatical patterning across discourse goes back at least as far as the early 1980's, in particular the work of Givón and his associates on measuring topicality and reference tracking (cf. the contributions in Givon (1994)). In Du Bois (1987), a pervasive pattern in the way given and new information is distributed across major syntactic functions was identified, which Du Bois labels Preferred Argument Structure. One of Du Bois' important contributions was to recognize that the traditional 'subject' category actually obscured an important difference between the way that intransitive subjects (S) and transitive subjects (A) behave with regard to the kinds of arguments that fill these functions. Thus, PREFERRED ARGUMENT STRUCTURE is formulated in terms of S, A and O (=transitive object). Two 'soft' constraints form the core of PAS:

(1) Avoid more than one lexical core argument per clause.
(2) Avoid lexical A's.

The first constraint simply states that, in natural discourse, a transitive clause (the only ones with two core argument positions, A and O) will very rarely have both argument positions filled by a lexical NP. At most, only one will be a lexical NP, while the other will be either zero (in languages which permit pronominal deletion), or pronominal. Du Bois (2003) cites figures from connected discourse in Hebrew, Sakapultek, Papago, English and Gooniyandi, none of which exhibit more than 7% of clauses with more than one core argument per clause.[1] The second constraint concerns which, if any, of the two arguments of a transitive clause will be the preferred host for a lexical NP. Here the tendency is quite clear, and has since been confirmed for a large number of languages: The A-role is strongly dispreferred as a host for lexical NPs, so if a transitive clause contains any lexical argument at all, it is generally in the O position.

What PAS is really all about is the way that given (hence generally non-lexical) and new (hence generally lexical) material is distributed across grammar. In theory, discourse status vis-á-vis the given/new distinction is logically quite independent of syntactic function. But in fact, there are significant and persistent patterns in the way the two interrelate. The tendencies noted by Du Bois can also be seen as descriptions of the manner in which new referents are typically introduced into discourse: they generally enter via the S, or the O function. In this respect then, PAS reveals a pervasive similarity linking S and O, while the A role patterns quite differently. Du Bois has referred to this as the 'discourse basis of ergativity', but as yet little progress has been made in demonstrating how the

[1] It is unclear from Du Bois' presentation whether the figures relate to the number of all clauses (including intransitives), or the total number of transitive clauses. The latter would obviously be the more relevant measure for assessing the validity of 'Avoid lexical A', because only transitive clauses could have more than one lexical argument anyway.

discourse patterns of argument realization should become entrenched in grammatical alignment patterns in morphosyntax.

Du Bois (1987) initially formulated his observations on the basis of data from Sakapultek, but they have proved to be remarkably robust generalizations, since confirmed in studies of very diverse languages, and across distinct text genres. The initial concentration on S, A and O has been broadened by the inclusion of I (Indirect Object), and some scholars have divided the S category into Se (equational clause), Sx (existential) and Si (other intransitive) (e.g. Clancy 2003). However, the core of PAS is still based on the interaction of the pragmatic dimension of givenness (in various gradations), and the syntactic functions of S, A and O.

Although PAS can now be hardly doubted as a statistical universal of connected discourse, there is little consensus regarding its explanation. Furthermore, a number of methodological and conceptual problems remain, in particular the undifferentiated use of the 'pronoun' category, the identification of the core roles, and differences across different grammatical persons. Some of these difficulties are taken up in Section 3 below where we discuss our own methodology.

Another line of typological research that relies on statistical methods is the investigation of REFERENTIAL DENSITY. While PAS refers to the distribution of lexical as opposed to pronominal and/or zero expressions of arguments, the term REFERENTIAL DENSITY refers to the ratio of overt expressions, lexical or pronominal, to all possible argument positions (Bickel 2003; Stoll & Bickel 2009):

(3)
$$RD = \frac{\text{N (number of overt arguments)}}{\textit{N (number of available argument positions)}}$$

Research by Balthasar Bickel and his associates has shown that language communities vary considerably in the extent to which speakers make use of overt NPs in discourse. Speakers of Chinese and Belhare (Sino-Tibetan), for instance, appear to be much more implicit about referents than speakers of Spanish or Russian who use overt NPs much more frequently (Bickel 2003:708; Stoll & Bickel 2009). Belhare exhibits a somewhat extreme case, as its speakers make extremely limited use of overt nominal expressions even in contexts where new referents are introduced into discourse.

While the statistical figures for RD are simply a matter of fact, the decisive factors behind it remain to be investigated cross-linguistically. It goes without saying that the RD of particular narratives may, for instance, vary according to different numbers of referents occurring in the story, even within the same language. But speakers of different languages produce considerably different values of RD even where they tell the same story. Although it is fairly clear that language communities have different narrative traditions, there are also some

indications that the typological profiles of languages may affect their varying RD values.

Bickel (2003) investigates the differences in RD between Pear Story narratives recorded in three languages of the Himalayas: Belhare (Sino-Tibetan), Nepali and Maithili (both IE) on the other. Despite their different genetic affiliations, the three linguistic communities are situated broadly in the same cultural context and make use of the same stock of traditional narrative conventions etc. Thus, a large set of factors very likely to influence the degree of RD are controlled for in this setup. The three languages differ, however, in their typological profile with regard to the role of case-marking in determining the controller in control constructions. In Belhare, overt morphological case plays but a marginal role in regulating control, which instead is sensitive to thematic relations. In the other two languages on the other hand, morphological case is crucial. Bickel finds evidence in support of a correlation between high RD on the one hand and the prominence of case-marking in determining control on the other.

In a somewhat similar vein, claims on the structural prerequisites for object pronoun deletion (so-called 'radical pro-drop') have been advanced by Neeleman & Szendröi (2007, 2008), but see Haig (2009) for criticism of the methodology and counter-evidence. But the point of these brief examples is that cross-linguistic differences that emerge through the quantitative comparison of discourse patterning may be rooted in deeper structural properties of the languages concerned, and it is surely part of the goal of typology to investigate this interaction further. Before further progress in this area can be made, there is a pressing need for extending the available typological database in spoken discourse.

3. METHODOLOGY: INTRODUCING GRAID

In this section we will briefly outline the main features of 'GRAID' ('Grammatical Relations and Animacy in Discourse'), the annotation system we propose. GRAID annotations are intended to facilitate research into PREFERRED ARGUMENT STRUCTURE, REFERENTIAL DENSITY and related topics in hitherto little-known languages. As the information necessary for investigating these topics cannot be derived in any way automatically, and detailed knowledge about a given language is required here, GRAID annotations have to be made by hand for different languages by linguists specializing in the respective languages.

GRAID adopts the basic principle of both Du Bois' and Bickel's proposals in that it relates the possible argument roles of each predicate with their actual realization in discourse. Thus GRAID-annotations always involve (i) analytical decisions on the argument structure of the predicate of a clause, and (ii) coding decisions on how a particular argument is to be classified. Both of these steps necessitate sound knowledge of the language concerned, and present the investigatgor with certain problems, some of which we discuss below. However,

these problems are not insurmountable, and are in principle no different to the problems faced by linguists providing morphosyntactic glosses.

Overt arguments are glossed for their formal type, i.e. lexical or pronominal, and their syntactic function S, A, P and a few other functions of arguments. For pronouns, we distinguish between free, weak, and bound forms. Unfilled argument positions are treated like zero arguments as opposed to lexical and pronominal ones, and receive a label for the syntactic function associated with this position.

The glossing for non-argument clause constituents is less elaborate. Predicates are labeled for their formal properties as either verbal, nominal or of other type, e.g. an adpositional phrase, but not for their semantics. Adjuncts are simply labeled for their adjunct status with no indication of their formal properties or semantics. Thus, in accordance with the research topic, the system is most elaborate in the domain of (core) arguments.

The RD and PAS of a GRAID-annotated text can then be determined on the basis of these annotations. For instance, we can easily search for all glosses of lexical and pronominal arguments and relate their number to the number of all glosses for lexical, pronominal, and zero expressions to determine the RD. PAS patterns are determined in a similar manner. In addition, GRAID also notes animacy properties of arguments, using a human vs. non-human distinction, but allowing for an intermediate stage for mythical beings endowed with human properties. Considerations of space prevent us from listing the full inventory of tags and instructions for their use. They are available online, together with examples. So far, we have worked on two languages, Gorani (West Iranian) and Vera'a (Austronesian, Oceanic), glossing a total of approx. 500 clause units. The system is workable and glossing becomes quite fast with a little practice. In the remains of this paper we will briefly outline a number of problems that we have encountered, and suggest solutions to them.

The first problem concerns identifying the clause units. Problems arise with serial verb or converb-type constructions, where a decision has to be made whether the verbs concerned have individual case frames (which would then increase the number of non-realized arguments in the gloss), or should be considered single predicates. Similar problems obtain with control-constructions after modals etc. A second, and perhaps more fundamental problem, concerns determining the potential argument structure of a verb. In head-marking languages with affixal cross-referencing of core arguments this may be relatively straightforward, but for the majority of languages, decisions must draw on the investigator's knowledge of the language concerned, and are inevitably to some extent arbitrary. Related to this problem is the question of distinguishing core from non-core arguments, again a perennial problem of any syntactic analysis that assumes this distinction. In particular in the realm of primary and secondary objects, addressees and other 'dative-type' arguments, investigators are required to make language-specific decisions. Finally, we should mention the issue of embedded complement clauses and relative clauses, which have on the one hand external properties as arguments and attributes, but have internal predicate-

argument structure. We are keenly aware of these difficulties, and have incorporated a number of annotational conventions into GRAID to facilitate these problems. However, we strongly wish to avoid inflating the inventory of tags beyond a practicable measure (approx. 30) in order to keep the system as simple as possible.

Obviously a certain proportion of any natural discourse will involve clause units which, for various reasons, cannot be meaningfully glossed (we consider approx. 10% an acceptable measure), and these are simply consigned to the "Non Classifiable" category. But our experience up to now has been that, for narrative monologues at least, GRAID annotations are applicable to vast majority of natural discourse.

REFERENCES

Bickel, B. 2003. Referential Density in Discourse and Syntactic Typology. *Language* 79(4), 708-736.

Clancy, P. 2003. The lexicon in interaction. Developmental origins of Preferred Argument Structure in Korean. In J.W. Du Bois et al. 2003, 81-108.

Du Bois, J.W. 1987. The Discourse Basis of Ergativity. *Language* 63(4), 805-855.

Du Bois, J.W. / I.F. Kumpf / W.J. Ashby (eds.). 2003. *Preferred Argument Structure: grammar as architecture for function.* Amsterdam: John Benjamins.

Givón, T. 1994. *Voice and Inversion.* Amsterdam: John Benjamins.

Haig, G. 2009. On the proposed correlation between discourse pro-drop and fusional pronominal case: Evidence from Iranian. Paper presented at the Third International Conference on Iranina Linguistics, Paris Sorbonne 3, September 11-13th, 2009.

Hofling, C. (2003): 'Tracking the deer. Nominal reference, parallelism and preferred argument structure in Itzaj Maya narrative genres'. in: Du Bois et al. 2003, 385-410.

Lamers, M./ S. Lestrade / P. de Swart (eds.) (2008): 'Animacy, Argument Structure, and Argument Encoding'. Special edition of *Lingua* 118, 156-218.

Neeleman, A. / Szendröi, K. 2007. Radical Pro-drop and the morphology of pronouns. *Linguistic Inquiry* 38(4), 671-714.

Neeleman, A. / Szendröi, K. 2008. Case morphology and radical pro-drop. In T. Biberauer (ed.). *The limits of syntactic variation*, 331-148. Amsterdam: Jon Benjamins.

Stoll, S. / B. Bickel (2009): 'How Deep are Differences in Referential Density'. in: Lieven, E. / J. Guo / N. Budwig / S. Ervin-Tripp / K. Nakamura / Ş. Özçalişkan (eds.): *Crosslinguistic Approaches to the Psychology of Language: Research in the Traditions of Dan Slobin*, 543–555. London: Psychology Press.