Relativization, Ergativity, and Corpus Frequency

Peter C. Gordon
University of North Carolina
Randall Hendrick
University of North Carolina

It is common to distinguish ergative languages from accusative languages. Whether this distinction is a surface morphological one or reflects a deeper division in how semantic composition is done has been the subject of debate (see Manning 1999 for a detailed overview). Ergative languages appear unusual because they are infrequent crosslinguistically. This view is corroborated by the fact that most ergative languages exhibit split ergative systems in which the ergative case-marking system holds in some structural contexts but the accusative emerges in others (Dixon 1979). In contrast, accusative languages typically lack ergative features. For this reason, some work has been done on how ergative languages could arise historically from accusative systems (see Trask 1979, Anderson 1976, Garrett 1990).

There is a tradition of thought in functional linguistics that offers a different view of the distinction between ergative and accusative languages. Fox (1987), for example, conjectures that ergativity is a highly natural feature of language and can be seen in common patterns of relativization. Specifically, Fox proposes the Absolutive Hypothesis (AH) in (1).

(1) Every language which has a strategy for relativizing must be able to relativize on at least S and P. (Fox 1987:864)

S abbreviates *subject of intransitive* and *P* abbreviates *patient*. The AH claims that relativization has the same pattern as ergative languages in the sense that subjects of intransitives and objects of transitives are case-marked absolutive and are, according to Fox, most easily relativized. This claim, Fox suggests, is directly at odds with Keenan and Comrie's (1977) relational hierarchy (RH). The RH holds that there is a hierarchy of grammatical relations as in (2) and that if a language relativizes on one point in this hierarchy, it relativizes on all points to its left.

(2) subject $<$ object $<$ oblique

The RH is organized around an accusative pattern in the sense that the term *subject* here identifies both the subject of a transitive clause and the subject of an intransitive clause. Fox's point is that the AH does not map directly onto a hierarchy like (2) because S and P do not share a common grammatical relation in that hierarchy. The AH is significant because it denies the view that ergative languages are unusual crosslinguistically. Indeed, Fox reports a study of relativization in an English spoken corpus that corroborates the AH.[1] Furthermore, it opens a very different view of the development of ergativity. Ergative languages emerge as the grammaticalization of semantic trends in discourse identified by Du Bois's (1987) Preferred Argument Structure Constraint, which hypothesizes, in part, that information

[1] There are, of course, other functional accounts that are consistent with the RH and that aim to explain corpus frequency, notably the account offered by Aissen (1999). In a more comprehensive work (Gordon and Hendrick 2004), we address the empirical viability of such approaches more generally.

Table 1
Frequency of extraction from relative clause in three corpora broken down by grammatical relations

	Extraction type		
Corpus	Subject	Object	Oblique
Brown	61.1% (1,606)	32.3% (848)	6.6% (174)
Switchboard	50.7% (506)	39.5% (394)	9.8% (98)
CHILDES	61.3% (473)	32.0% (247)	6.7% (52)

packaging in conversation is skewed along an ergative basis. In this sense, the AH expresses pressures within language use as social interaction and is contrasted by Fox with the RH, which is typically linked to claims about individuals' cognition.

We investigated the frequency of different kinds of relative clauses (RCs) in three different English corpora: the Brown corpus (Kučera and Francis 1967), the Switchboard corpus (Godfrey, Holliman, and McDaniel 1992), and the CHILDES corpus (MacWhinney 2000). We tabulated features of 2,628 RCs in randomly sampled tokens from Brown and 998 randomly selected tokens from Switchboard. Because the CHILDES corpus is not parsed, we reviewed all of the utterances containing *that* produced by children age 5 or younger in any of the US samples meeting MacWhinney's (2000) guidelines for coding conversations and found those instances in which *that* functioned as a complementizer that introduced an RC. This method yielded 772 utterances containing RCs. Criteria were established for judging characteristics of interest about the RCs. Three coders were instructed on these criteria, and the reliability of their judgments was established by having them independently judge the same randomly sampled subsets of the data. All pairs of coders achieved interrater reliability (kappa $>$ .8; Carletta 1996, Siegel and Castellan 1988) on each type of judgment. After reliability was established, the coders worked individually, with each of them coding about one-third of the data. The evidence obtained in this way will be used first to evaluate the RH and then to evaluate the AH.

Table 1 shows the frequency in the three corpora of the types of RCs specified by the RH (subject, object, and oblique extractions) that were coded. For all three corpora, subject extractions occurred more frequently than object extractions, which in turn occurred more frequently than oblique extractions, thereby showing that the relative frequency of types of RCs within English matches the crosslinguistic pattern captured by the RH.[2]

[2] The statistical significance of differences in frequencies for pairs of extraction types (i.e., subject vs. object and object vs. oblique) was assessed using the normal approximation to the binomial. For table 1, all such comparisons were highly significant ($p < .0001$) for all three corpora.

The patterns of relative frequency are consistent across the corpora despite many pressures that might make the corpora different from one another; ''genre effects'' are well documented for many linguistic phenomena (Biber 1988). We will not attempt to explain apparent variation in the magnitude of differences in the corpora but will restrict our attention to patterns of relative frequencies. For example, the proportions of different types of extractions shown in table 1 for Brown and CHILDES are close but Switchboard has a smaller proportion of subject extractions. The reason for this apparent difference in magnitude is unclear, but there is a consistent ordering for relative frequency in all three corpora, a finding that is not consistent with Fox's (1987) claim that frequency evidence in support of the RH comes only from written corpora of English.

This pattern of relative frequency is also inconsistent with Fox's (1987) report that her spoken corpus contains an equal proportion of subject-extracted and object-extracted relatives. Keenan (1975), in a study of written text, reported a higher proportion of subject-extracted relatives than object-extracted relatives, in accord with the RH. Fox discounts this evidence on the grounds that corpora of written language are less influenced by information-packaging pressures than are corpora based on naturally occurring conversation.[3] Our results confirm Keenan's positive findings in a larger corpus of written text and also extend that finding to two corpora that used very different methods to sample spoken conversations.

Table 2 shows the frequency in the three corpora of the types of RCs specified by the AH (subject of transitive *A*, subject of intransitive *S*, and object *P*).[4] The AH fails to describe the relative frequency of extraction across all three corpora. For the Brown corpus, A extraction is most common, a pattern opposite to that expected under the AH. In contrast, for Switchboard and CHILDES, A extraction is the least common, a pattern expected under the AH.[5] That the AH finds support in data from these two corpora is consistent with Fox's focus on spoken language. In addition, it shows that from the perspective of frequency data, the RH and AH are not mutually exclusive (as Fox presents them). However, a comparison of the patterns in tables 1 and 2 gives two reasons to prefer the RH to the AH. First, the RH successfully predicts the relative frequency of extraction types in all three corpora, while success for the AH requires that one of the corpora be discounted. Second, the RH correctly specifies the relative frequency of the three

[3] Specifically, written texts are less influenced by demands of ''anchoring,'' which require utterances to be cohesive and which encourage interlocutors to attend to one another (Fox 1987:861n10).

[4] The transitive versus intransitive breakdown for oblique extractions is as follows: Brown (106 vs. 68), Switchboard (32 vs. 66), and CHILDES (16 vs. 36). This breakdown is not directly relevant to the AH, but we include the information for the sake of completeness.

[5] These differences were again assessed using the normal approximation to the binomial. Pairwise comparison of A with S and of A with P was highly significant in all three corpora ($p < .002$ in all cases).

Table 2
Frequency of extraction from relative clause in each of the three corpora classed by whether the extraction site is subject of transitive, subject of intransitive, or object. The percentage of NPs in each cell that are definite is shown in parentheses. If the definiteness status of the NP could not be determined for some cases contributing to the count, then the number on which the percentage is based is also given. (This occurred if the sentence was ungrammatical or incomplete. The vast majority of these cases occurred in speech by children (CHILDES).)

	Extraction type		
Corpus	Subject of transitive (A)	Subject of intransitive (S)	Object of transitive (P)
Brown	41.5% (1,016) *(48.3% definite)*	24.1% (590) *(44.7% definite)*	34.4% (841) *(61.9% of 839 definite)*
Switchboard	24.5% (220) *(30.9% definite)*	31.8% (286) *(39.9% definite)*	43.7% (393) *(50.4% definite)*
CHILDES	27.7% (198) *(44.3% of 176 definite)*	37.9% (271) *(53.6% of 265 definite)*	34.4% (246) *(53.3% definite)*

types of extractions that it considers, while the AH specifies the relative frequency of A with respect to S and P but does not specify the relative frequency of S and P with respect to each other. In fact, for the two corpora in which the AH is successful, the relative frequency of S and P differs.[6] To summarize this comparison: the RH is both more specific in its predictions and more generally successful than the AH in accounting for these corpus data.

Fox suggests that the AH is natural because it embodies Du Bois's (1987) Preferred Argument Structure Constraint (PAS). The PAS is designed to characterize how information is packaged in a discourse. On this view, information packaging has an ergative/absolutive foundation because the PAS holds that new discourse entities will be more likely to appear in S or P rather than A, exactly the positions that are assigned absolutive case in ergative/absolutive languages. In addition, the PAS recognizes a preference to use only one semantic argument per unit of discourse. On the common assumption that definite NPs

[6] For Switchboard, the greater frequency of P as compared with S is highly significant ($p < .000001$) as tested by the normal approximation to the binomial. For CHILDES, the greater frequency of S as compared with P was not reliable ($p > .2$). The association between corpus (Switchboard vs. CHILDES) and grammatical role (S vs. P) was highly reliable ($\chi^2 = 11.1$, $p < .001$).

Table 3
Frequency of NPs in different syntactic roles when modified by relative clauses

	Modified NP		
Corpus	Subject	Object	Oblique
Brown	14.9% (391)	35.1% (924)	50.0% (1,313)
Switchboard	16.8% (168)	52.6% (525)	30.6% (305)
CHILDES	12.9% (95)	56.5% (418)	30.6% (226)

encode old, familiar information more naturally than they encode new information (Lyons 1999, Prince 1981, Comrie 1989, among many others), the PAS leads us to expect that P and S should be less definite than A. Fox (1987:860) follows a similar train of thought to predict the frequency of pronouns, which also encode old information.[7] Table 3 shows the percentage of each type of NP that is definite. In no instance does A provide the highest percentage of definite NPs; indeed, definite NPs tend to occur in S or P position.[8] This contradicts the prediction that we derived from the PAS and AH that because S and P are thought to present new information, they should be less likely to contain definite NPs than should A.

An alternative explanation for why A is less common as an extraction site in the spoken corpora but not in Brown might build on the results of table 3. The two spoken corpora contained more RCs whose heads functioned as direct objects than RCs whose heads functioned as oblique heads, while the written corpus contained more RCs whose heads functioned as obliques.[9] We hypothesize that this difference emerged from the differing pressures in contexts of speaking versus writing. A sentence that contains an oblique modified by an RC will tend to have more constituents than the other two types of sentences. Because writing involves less immediate time pressure than speaking, more cognitive resources for producing such an elaborate sentence may be available in writing. This cognitive conjecture can be extended to the differences between the corpora shown in table 2. The A and S extractions shown in table 2 represent subject extractions from transi-

[7] Fox (1987:864) also examines an empirical prediction concerning the frequency of definite Ps in RC as compared with their frequency in all utterances. Our data do not allow us to evaluate this claim because we did not code all Ps in the three corpora.

[8] All corpora show significant associations between definiteness and A versus S or P: Brown ($\chi^2 = 9.9$, $p < .002$), Switchboard ($\chi^2 = 15.4$, $p < .001$), and CHILDES ($\chi^2 = 4.3$, $p < .05$).

[9] All differences are significant ($p < .0001$) by the normal approximation to the binomial.

tive verbs and intransitive verbs, respectively. Because transitive verbs have more arguments than intransitive verbs, the contrasting relative frequencies between A and S in Brown versus Switchboard and CHILDES could simply reflect the same cognitive pressures that lead to greater modification of obliques in Brown as compared with Switchboard and CHILDES. The diminished time pressure in writing as compared with speaking could facilitate the production of more complex sentences with more arguments.

This cognitive explanation—that resource limitations tend to cause a reduction in the number of arguments in a sentence—accounts for differences between spoken and written language with respect to the incidence of extraction from the subject position of transitive verbs (table 2) and the kinds of NPs that are modified by RCs (table 3). It provides an alternative to the PAS, which claims to be a constraint on language use as a social activity and which does not appear to offer a ready explanation of why spoken and written language would differ with respect to the type of NP that is modified by an RC. Cognitive explanations of the relative frequency of different types of extraction have a long history (Keenan 1975), and psycholinguistic evidence shows that object-extracted RCs are more difficult to comprehend than subject-extracted RCs (Caramazza and Zurif 1976, Just et al. 1996, MacWhinney 1982).

Our enthusiasm for this cognitive explanation is tempered by two considerations. First, while it is plausible that resource limitations cause a reduction in the number of arguments in a sentence, we do not know of any psycholinguistic evidence that specifically supports that idea. Second, there are clear instances where the difficulty of understanding sentences with RCs does not pattern with strong differences in the frequency of syntactic and semantic characteristics of RCs (Gordon, Hendrick, and Johnson 2004). This negative finding raises the *grain* problem, which has caused substantial difficulty for efforts to relate language-use statistics to language processing (Mitchell et al. 1995, Townsend and Bever 2001). It is always possible to argue that the critical relationship between use statistics and processing ease occurs at a coarser or finer grain of analysis than was examined in any particular study. No general theory has yet been developed that specifies the critical level of analysis that relates use to processing ease.

The grain problem also exists for efforts to employ language-use statistics to study language structure. For the present data, the patterns predicted by the RH are found in all three corpora despite the many differences among them (i.e., written vs. spoken language, adult vs. child speech, face-to-face communication vs. phone conversations). The structural principles captured by the RH are sufficiently powerful that they emerge in language-use data of English despite these many differences that are incidental to language structure. We believe that such consistency of patterns across communicative situations should be the basis for determining whether language-use data should be used to evaluate theoretical claims about language structure.

References

Aissen, Judith. 1999. Markedness and subject choice in Optimality Theory. *Natural Language & Linguistic Theory* 17:673–711.

Anderson, Stephen R. 1976. On the notion of subject in ergative languages. In *Subject and topic*, ed. by Charles Li, 1–23. New York: Academic Press.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Caramazza, Alfonso, and Edgar Zurif. 1976. Dissociation of algorithmic and heuristic processes in sentence comprehension: Evidence from aphasia. *Brain and Language* 3:572–582.

Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics* 22:249–254.

Comrie, Bernard. 1989. *Language universals and linguistic typology: Syntax and morphology*. 2nd ed. Chicago: University of Chicago Press.

Dixon, R. M. W. 1979. Ergativity. *Language* 55:59–138.

Du Bois, John. 1987. The discourse basis of ergativity. *Language* 63: 805–855.

Fox, Barbara. 1987. The noun phrase accessibility hierarchy revisited. *Language* 63:856–870.

Garrett, Andrew. 1990. The origin of NP split ergativity. *Language* 66:261–296.

Godfrey, John, Edward Holliman, and Jane McDaniel. 1992. Telephone speech corpus for research and development. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. Piscataway, N.J.: IEEE.

Gordon, Peter C., and Randall Hendrick. 2004. Relative clauses, frequency, and dimensions of markedness. Ms., University of North Carolina at Chapel Hill.

Gordon, Peter C., Randall Hendrick, and Marcus Johnson. 2004. Effects of noun phrase type on sentence complexity. *Journal of Memory and Language* 51:97–114.

Just, Marcel A., Patricia A. Carpenter, Timothy A. Keller, William F. Eddy, and Keith R. Thulborn. 1996. Brain activation modulated by sentence comprehension. *Science* 274:114–116.

Keenan, Edward. 1975. Variation in Universal Grammar. In *Analyzing variation in language*, ed. by Ralph Fasold and Roger Shuy, 138–148. Washington, D.C.: Georgetown University Press.

Keenan, Edward, and Bernard Comrie. 1977. Noun phrase accessibility and Universal Grammar. *Linguistic Inquiry* 8:63–99.

Kučera, H., and W. N. Francis. 1967. *Computational analysis of present-day American English*. Providence, R.I.: Brown University Press.

Lyons, Christopher. 1999. *Definiteness*. Cambridge: Cambridge University Press.

MacWhinney, Brian. 1982. Basic syntactic processes. In *Language*

acquisition. Vol. 1, *Syntax and semantics*, ed. by Stan Kuczaj, 73–136. Hillsdale, N.J.: Lawrence Erlbaum.
MacWhinney, Brian. 2000. *The CHILDES project: Tools for analyzing talk*. 3rd ed. Mahwah, N.J.: Lawrence Erlbaum.
Manning, Christopher. 1999. *Ergativity: Argument structure and grammatical relations*. Stanford, Calif.: CSLI Publications.
Mitchell, Don C., Fernando Cuetos, Martin M. B. Corley, and Marc Brysbaert. 1995. Exposure-based models of human parsing: Evidence for the use of coarse-grained (nonlexical) statistical records. *Journal of Psycholinguistic Research* 24:469–488.
Prince, Ellen. 1981. Towards a taxonomy of given-new information. In *Radical pragmatics*, ed. by Peter Cole, 223–255. New York: Academic Press.
Siegel, Sidney, and N. John Castellan. 1988. *Nonparametric statistics for the behavioral sciences*. 2nd ed. New York: McGraw-Hill.
Townsend, David J., and Thomas G. Bever. 2001. *Sentence comprehension: The integration of habits and rules*. Cambridge, Mass.: MIT Press.
Trask, Robert L. 1979. On the origins of ergativity. In *Ergativity: Towards a theory of grammatical relations*, ed. by Frans Plank, 385–404. New York: Academic Press.

The Structure of Children's Linguistic Knowledge

Andrea Gualmini
University of Maryland
Stephen Crain
University of Maryland

A recurring theme in arguments from the poverty of the stimulus concerns children's knowledge of linguistic structure. Nativists point to the extensive gap between what children know and what they could have learned from experience, even given optimistic assumptions about children's abilities to extract information from the environment, and to form generalizations. This squib looks at children's knowledge of linguistic structures that involve the semantic property of downward entailment, allowing us to address a recent critique of children's knowledge of structure offered by Lewis and Elman (2002).

1 Structure Dependence and Poverty of the Stimulus

An example of structure-dependent linguistic principles deals with question formation. This phenomenon was originally described by Chomsky (1971), who questioned the extent to which the primary linguistic data could lead children to form the correct generalizations relating declarative sentences and their yes/no question counterparts (see also Chomsky 1980 and discussion in Piattelli-Palmarini 1980).

We thank Amanda Gardner, Beth Rabbin, and Nadia Shihab for their help in conducting the experiments, and Norbert Hornstein, Luisa Meroni, Paul Pietroski, and Rosalind Thornton for helpful discussion. We also thank the staff, teachers, and children at the Center for Young Children at the University of Maryland at College Park, where the experiments were conducted.