SEEKING COMPOSITIONALITY IN HOLISTIC PROTO-LANGUAGE WITHOUT SUBSTRUCTURE – DO COUNTEREXAMPLES OVERWHELM THE FRACTIONATION PROCESS?

SVERKER JOHANSSON

School of Education and Communication, University of Jönköping, Box 1026
SE-551 11 Jönköping, Sweden
lsj@hlk.hj.se

In holistic theories of protolanguage, a vital step is the fractionation process where holistic utterances are broken down into segments, and segments associated with semantic components. One problem for this process may be the occurrence of counterexamples to any segment-meaning connection. The actual abundance of such counterexamples is a contentious issue (Smith, 2006; Tallerman, 2007). Here I present calculations of the prevalence of counterexamples in model languages. It is found that counterexamples are indeed abundant, much more numerous than positive examples for any plausible holistic language.

1. Introduction

Human beings today have language. Our ancestors long ago did not. The notion that modern language with all its complexity arose *ex nihilo* is preposterously unlikely, which implies that one or more intermediate stages, less complex than modern language, must have existed. A popular possibility for an early intermediate stage is a language where each utterance is a unit without substructure. In analogy with the ontogeny of language, we might call this a one-word stage.

There are at least two ways to get from a one-word stage to a composite language, either analytic/holistic or synthetic (Hurford, 2000; Bickerton, 2003). In the holistic version (Wray, 2000; Arbib, 2003), the units of the one-word stage are holistic utterances, which are then fractionated into parts that become independent recombinable morphemes in the next stage, whereas in the synthetic version (Bickerton, 2000; Jackendoff, 2002, among others), two or more units from the one-word stage are combined into structured utterances in the next stage.

The segmentation and analysis step, finding substructure in utterances that are postulated to lack substructure, is a critical step for holistic theories. It is not obvious to me, nor to Bickerton (2003) or Tallerman (2007), why the fractionation process envisaged by Wray (2000) would be expected to work. A similar process is certainly present in modern-day language acquisition — children first acquire some stock phrases as unanalyzed wholes, and later figure out their internal struc-

ture — but that works only because these stock phrases *have* an internal structure, given by the grammar of the adults from whom the child acquires them. As an analogy for the origin of grammar, this is unsatisfactory.

Wray (2000) describes a scenario in which people already talking at the one-word stage at some point acquire a grammar from somewhere — apparently not from any linguistic or communicative pressures, but as an exaptation — and start applying it to their language, attempting to identify structure and constituents in their structureless holistic one-word utterances.

Tallerman (2004, 2007) provides a detailed critique of this process, to which Smith (2006) provides a partial response. In this paper, I will concentrate on one specific point of contention between Tallerman and Smith, which concerns how connections are established between semantic components and sound segments.

By pure chance, it may sometimes happen that different utterances have both a phonetic segment in common, and a semantic component in common. It is argued by e.g. Wray (2000) that this will lead to the identification of the phonetic segment with the semantic component, so that the former comes to "mean" the latter. Tallerman (2004, 2007) argues that it is self-evident that counterexamples will by far outnumber confirming examples for such a generalization. Smith (2006) disagrees, arguing that there is no logical necessity that counterexamples outnumber positive examples.

Smith (2006) further argues that it is not established that counterexamples, whatever their frequency, are actually fatal to generalization. This issue hinges on whether the analysis process in proto-humans has a logical and statistical component, or is purely based on positive examples. The mental processes of proto-humans are unfortunately unavailable to direct observation, but since it has been established that both modern human infants (Saffran et al, 1996) and monkeys (Hauser, Newport, & Aslin, 2001) are sensitive to statistical patterns in language-like input, it is not parsimonious to assume that proto-humans totally disregard statistics.

The weight of Tallerman's argument thus depends on the actual ratio of counterexamples to positive examples in plausible proto-languages, a ratio that can be estimated through simple calculation in simulated model languages. I present here the results of such a calculation.

2. Model

A toy language is constructed by creating a set of utterances. Each utterance consists of a number of phonological segments, and carries a meaning consisting of a basic predicate-argument structure, with a single predicate and one or more arguments. Both phonological segments and meaning are randomly assigned to each utterance, uncorrelated with each other.

The following features of the language could be varied as free parameters in the model:

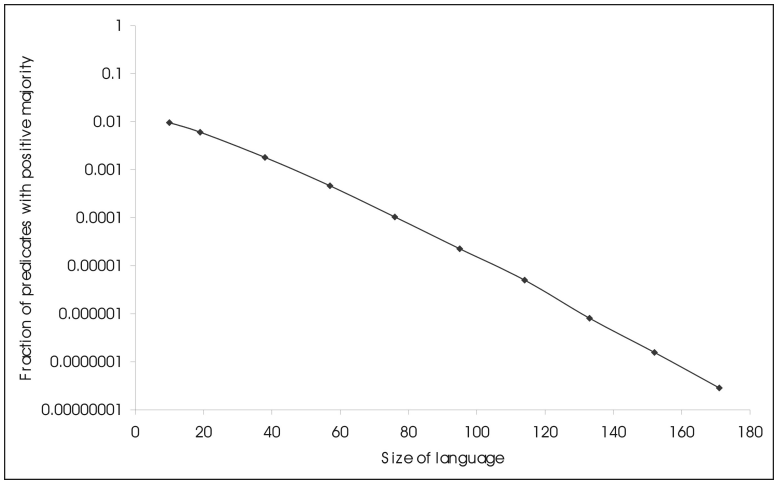

Figure 1. The fraction of predicates for which positive examples outweigh counterexamples, as a function of the size of the language. The values of the other parameters are fixed at #segments = #predicates = 50, utterance length = 4 segments.

- Total size of language, number of distinct holistic utterances
- Total inventory of phonological segments
- Total semantic inventory of predicates
- Total semantic inventory of arguments
- Number of phonological segments in one utterance

Many different parameter combinations were investigated, to identify which regions, if any, in parameter space are conducive to creating a composite language as argued by Wray (2000) and Smith (2006). For each parameter combination, a large number of toy languages (100,000 or more) were generated and analysed.

Once a language has been randomly generated with a given set of parameters, it is analysed for possible semantic-phonological connections according to the following procedure:

- For all predicates and all phonological segments in the language, the number of co-oocurrences of predicate p_i with segment s_j in the same utterance are counted.
- For each predicate in the language, the phonological segment s_{best} that most often co-oocurs with it is identified.

- For the segment s_{best}, both the number of positive examples, where it co-occurs with p_i, and the number of counterexamples occurring anywhere in the language, are counted. A counterexample can be either the occurrence of s_{best} in an utterance that does not mean p_i, or an utterance that means p_i but does not contain s_{best}.

Various higher-order complications, like the possibility that the same segment s is the best choice for two different predicates, have been neglected. Taking such complications into account would only decrease the possibility of finding and reinforcing connections. It is also assumed for the sake of the calculation here, *contra* Tallerman (2007), that segmentation of an utterance is unproblematic, and that proto-humans already have compositional semantics.

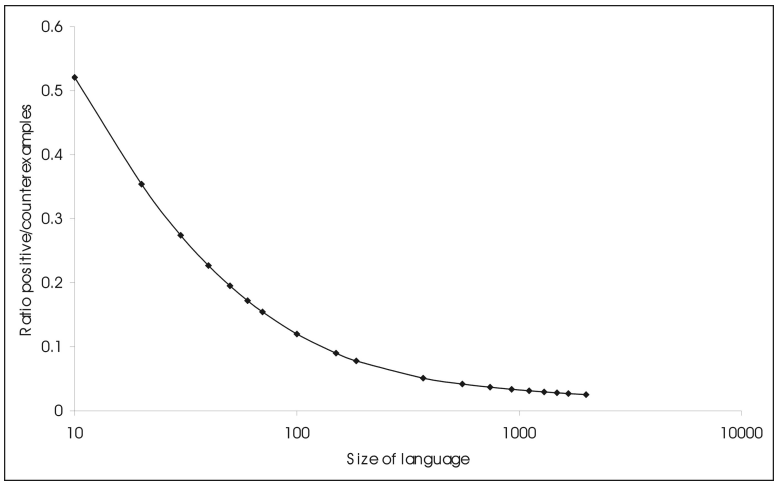

Figure 2. The ratio of positive examples to counterexamples, as a function of the size of the language. The values of the other parameters are fixed at #segments = #predicates = 50, utterance length = 4 segments.

3. Results

For all parameter combinations, the number of counterexamples were found to outweigh the number of positive examples by a considerable margin. For no parameter combination did the fraction of all predicates with more positive examples than counterexamples exceed 2% (Fig. 1).

The most important parameter is language size. The smaller the language, the larger the fraction of predicate-segment connections with predominantly positive examples, as shown in Fig. 1, and the larger (but still much less than unity) is

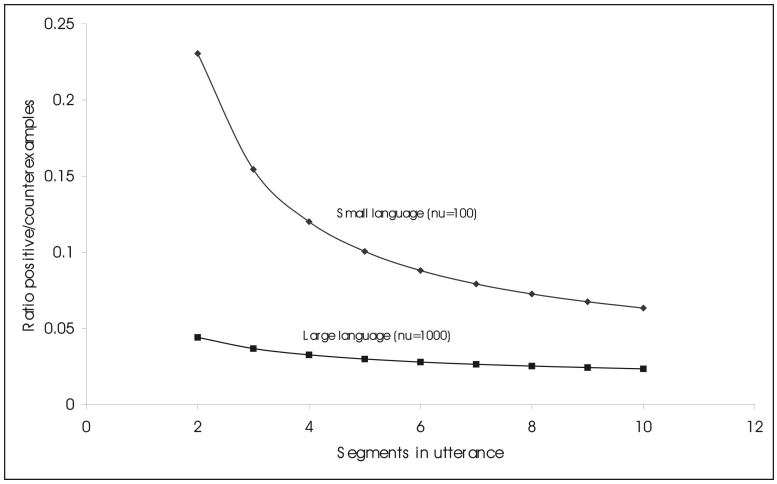

Figure 3. The ratio of positive examples to counterexamples, as a function of utterance length, for two different language sizes. The values of the other parameters are fixed at #segments = #predicates = 50.

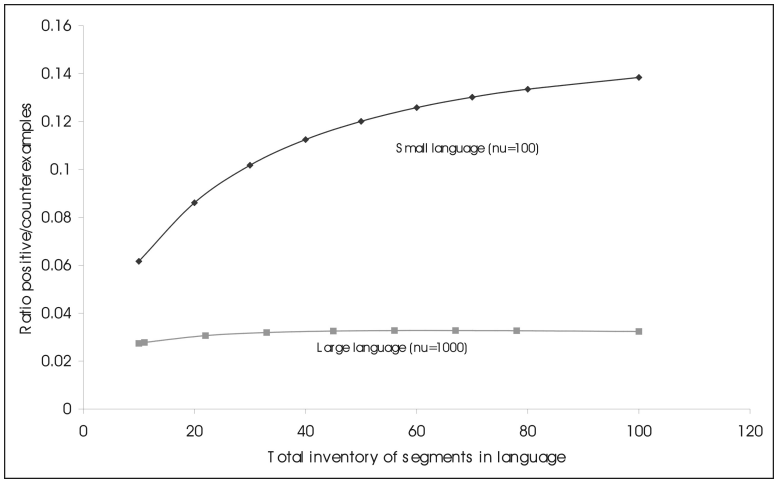

Figure 4. The ratio of positive examples to counterexamples, as a function of segment inventory, for two different language sizes. The values of the other parameters are fixed at #predicates = 50, utterance length = 4 segments.

the ratio of positive examples to counterexamples. This can be explained as a sampling effect, with random fluctuations being more important at small sample

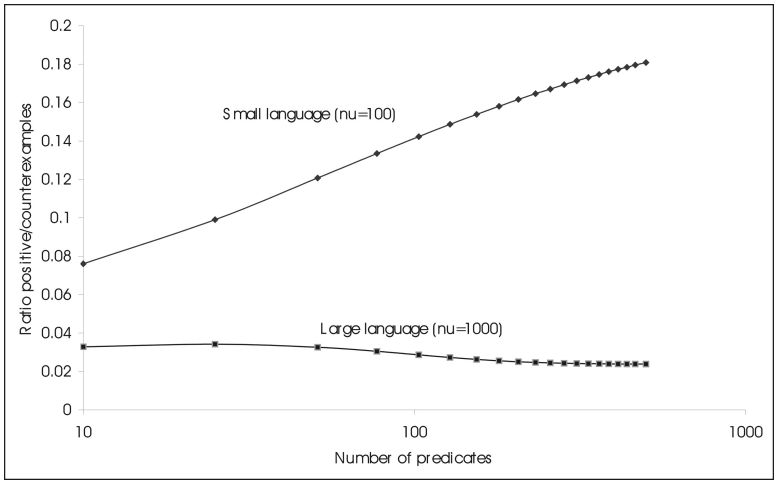

Figure 5. The ratio of positive examples to counterexamples, as a function of predicate inventory, for two different language sizes. The values of the other parameters are fixed at #segments = 50, utterance length = 4 segments.

size.

Similarly, the number of segments per utterance has a substantial effect, with very short utterances being "better", as shown in Fig. 3.

For small languages the connection success gradually grows with increasing segment inventory and predicate inventory (Figs. 4 and 5, upper curves). For large languages, the situation is different. Success rate is very low, largely independent of both segment inventory and predicate inventory (Figs. 4 and 5, lower curves).

4. Discussion

It is clear that there is only a small range of parameters for which the positive examples are not totally overwhelmed by counterexamples. The fractionation process has a non-negligible chance of success only for very small simple languages – but where would the pressure towards compositionality come from with a tiny language? And even for these tiny languages, success rate is small unless the inventory of segments and predicates is of the same order of magnitude as the total number of utterances in the language, which is hardly plausible. Unless it can be shown that humans totally disregard counterexamples when extracting patterns from data, the argument from counterexamples has considerable force.

References

Arbib, M. A. (2003). The evolving mirror system: a neural basis for language readiness. In M. H. Christiansen & S. Kirby (Eds.), *Language evolution*. Oxford: Oxford University Press.

Bickerton, D. (2000). How protolanguage became language. In Knight, Studdert-Kennedy, & Hurford (Eds.), *The evolutionary emergence of language*. Cambridge: Cambridge University Press.

Bickerton, D. (2003). Symbol and structure: a comprehensive framework. In M. H. Christiansen & S. Kirby (Eds.), *Language evolution*. Oxford: Oxford University Press.

Hauser, Newport, & Aslin. (2001). Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. *Cognition 78:B53-B64*.

Hurford, J. R. (2000). Introduction: the emergence of syntax. In Knight, Studdert-Kennedy, & Hurford (Eds.), *The evolutionary emergence of language*. Cambridge: Cambridge University Press.

Jackendoff, R. (2002). *Foundations of language. brain, meaning, grammar, evolution*. Oxford: Oxford University Press.

Saffran et al. (1996). Statistical learning by 8-month old infants. *Science 274:1926-1928*.

Smith, K. (2006). The protolanguage debate: bridging the gap. In A. Cangelosi, A. Smith, & K. Smith (Eds.), *The evolution of language proceedings of the 6th international conference (evolang6) rome, italy 12 - 15 april 2006*. Singapore: World Scientific Publishing.

Tallerman, M. (2004). Analysing the analytic: problems with holistic theories of the evolution of protolanguage. In *Proceedings of 5th conference on evolution of language, leipzig*.

Tallerman, M. (2007). Did our ancestors speak a holistic protolanguage? *Lingua 117:579-604*.

Wray, A. (2000). Holistic utterances in protolanguage: the link from primates to humans. In Knight, Studdert-Kennedy, & Hurford (Eds.), *The evolutionary emergence of language*. Cambridge: Cambridge University Press.