Kolmogorov Complexity in Language

Denis Paperno

This paper argues the merits of applying the computational (algorithmic) notion of Kolmogorov complexity to linguistic complexity. Instead of discussing complexity of particular syntactic structures [1], I propose to consider the complexity of language itself as measured by the complexity of its grammar. Two limits on the utility of Kolmogorov complexity will be noted, but despite them I support instances where Kolmogorov complexity yields interesting result in grammatical theory.

Kolmogorov complexity of an object is defined as the length of the shortest description of that object (usually, objects are represented as strings and their descriptions as programs that produce those strings as an output) [3]. Descriptions of (formal) languages are grammars, and Kolmogorov complexity of a language is the length of its shortest grammar. This corresponds well to the pre-theoretical notion of complexity: if a language L has lots of exceptions and idiosyncrasies, its grammar needs to account for these, and will be longer than a grammar for a language L' that lacks such irregularities. Therefore L' is simpler than L in the sense defined above. Other things being equal, languages with elaborate agreement systems like Basque would be more complex than languages like Mandarin without any agreement because their grammar would require either more rules or longer rules to account for agreement patterns. If a language L has a richer lexicon than L', the description of L has to account for this by adding more vocabulary items, so L is, once again, more complex.

One fundamental property of Kolmogorov complexity is that it depends on the choice of metalanguage ("description method"). This tells us that *complexity is inherently theory-dependent.* For example, assume a Principles and Parameters model where all crosslinguistic differences are represented by values of a fixed finite set of parameters. If all parameters are binary, a description of a language is just a sequence of 0s and 1s encoding values of individual parameters. All such descriptions have the same length (= number of parameters) and therefore all languages have the same complexity. If we don't adhere to strict P&P but use, say, phrase structure rules as a description method, then languages may in principle differ in complexity.

Is there a way to measure the complexity of a language? One of the results of the algorithmic complexity theory is that there is no algorithm for measuring Kolmogorov complexity (of an arbitrary finite object), there are only approximate estimates. Since (formal) languages include all finite objects (singleton languages), there is no algorithm for measuring the complexity of an arbitrary language. This means that it may not always be possible to determine with confidence if one language is simpler than another.

Another fact that can be useful for us linguists is that complexity is independent of expressive power. It is best to illustrate this thesis with mathematical languages whose expressiveness is well understood. For example, standard propositional logic with five logical operators (1) is more complex than logic with one operator (2). The two logics are known to be expressively equivalent. A converse example comes from predicate logic: if the universal quantifier $\forall$ is replaced by the Rescher quantifier Q^R, structural properties of the logical language, including its complexity, remain the same. Yet the expressiveness of the logic with Q^R ($\approx$'most things') is greater than that of standard first order logic [4, 474].

The two examples just discussed come from logic. Can one find analogous crosslinguistic differences in either complexity or expressiveness in natural language? A recent crosslinguistic survey of quantification in natural language [2] provides an appropriate testing ground for this issue. The study found that expression of logical operators on quantifiers such as conjunction and disjunction differ crosslinguistically in ways that suggest a difference in complexity (cf. (3) and its English translation). But the study didn't reveal differences in expressivity. For example, all languages considered can express proportional quantifiers like *most.* It is known that first order logic (one of my artificial examples) can not express such quantifier concepts.

Grammatical complexity is an extremely elusive property: while being an intuitively simple and adequate notion, it can not be measured, and is highly theory-dependent. But Kolmogorov complexity allows us to understand this elusiveness precisely. The definition of Kolmogorov complexity predicts that in the Principles and Parameters theory all languages have equal complexity, while other theories may allow typological variation in complexity. Theory of Kolmogorov complexity proves as a theorem that complexity can not be algorithmically measured. Last but not least, it allows us to dissociate complexity from expressivity as distinct properties of languages.

(1) Propositional Calculus

1.$S \rightarrow (SRS)$

2.$R \rightarrow \wedge$

3.$R \rightarrow \vee$

4.$R \rightarrow \Rightarrow$

5.$R \rightarrow \Leftrightarrow$

6.$S \rightarrow \neg S$

7.$S \rightarrow P$

8.$P \rightarrow P1$

9.$P \rightarrow p$

(2) Propositional logic with Pierce's arrow

1.$S \rightarrow (S \downarrow S)$

2.$S \rightarrow P$

3.$P \rightarrow P1$

4.$P \rightarrow p$

(3) Awa wala Ayda, am-na k-u ci wey
awa or ayda exist-FIN CL-C$_{\text{REL}}$ PART sing
'Either Awa or Ayda sang'
(Lit. 'Awa or Ayda, there is someone among them who sang') [5]

References

[1] Robin Clark. Kolmogorov complexity and the information content of parameters. Technical report, Institute of Research in Cognitive Science, University of Pennsylvania, October 1994.

[2] Edward L. Keenan and Denis Paperno, editors. *Handbook of Quantifiers in Natural Language*. Springer, to appear.

[3] Paul Vitanyi Ming Li. *An Introduction to Kolmogorov Complexity and its Applications*. Springer, 2nd edition, 1997.

[4] Stanley Peters and Dag Westerståhl. *Quantifiers in Language and Logic*. Clarendon Press, Oxford, 2006.

[5] Khady Tamba, Harold Torrence, and Malte Zimmermann. Wolof quantifiers. In Edward L. Keenan and Denis Paperno, editors, *Handbook of Quantifiers in Natural Language*. Springer, to appear.