DETAILS MATTER: PROBLEMS AND POSSIBILITIES FOR MEASURING CROSS-LINGUISTIC COMPLEXITY

DANIEL ROSS[*1]

[*]Corresponding Author: djross3@gmail.com
[1]Department of Linguistics, University of Illinois at Urbana-Champaign, USA

1. Purpose

Expanding on the theoretical proposal in Ross (2014), I test the implications and feasibility of a detail-oriented, frequency-independent metric of complexity as applied to a large sample of languages, from the perspective of linguistic typology. I measure *effective complexity* in the sense of Gell-Mann (1994, p. 58), characterizing the complexity of a system as the number of systematic rules required to describe it (e.g., grammar) while removing stochastic information (e.g., vocabulary). We can imagine the complexity of a language as the length of an ideal descriptive grammar: more paragraphs for more complex languages.[1] The task at hand is to implement such a metric for 37 languages from the Universal Dependencies project (http://universaldependencies.org/), as provided by the workshop organizers.[2] This presents a unique challenge for Ross's proposal, which asserts that accurately measuring complexity requires an exhaustive description of a language. But can we *estimate* linguistic complexity?

2. Estimating complexity from corpus data

How many rules are there in a language? Theoretical perspectives on the subject vary widely. Chomsky's (1995) Minimalist Program strives to reduce all of the possible rules from earlier syntactic theories to the minimum number required to

[1] And indeed the length of the paragraphs themselves, indicating the relative complexity of each feature. But that can only be measured after identifying all relevant features in the first place. For a literal answer to this question of length of descriptive grammars, see Section 5.

[2] A purely syntax-based measure is proposed based on the provided corpus data, although a full measure of complexity would also include other features (morphology, phonology, etc.).

explain the data. In the extreme, there may be only one rule of core syntax (*Merge*, combining two elements to form a larger phrase), but additional (possibly language-specific, or interface-based) rules beyond core syntax are required to explain the full range of cross-linguistic variation. At the other extreme, Construction Grammar (Goldberg, 1995, *inter alia*) posits an indefinite number of syntactic constructions, presumably stored like vocabulary in the lexicon. That introduces another problem for measuring complexity: if syntactic constructions are arbitrary like lexical items, perhaps they should be considered stochastic information and disregarded from our measurements of *effective complexity*. Regardless, we can reasonably assume that any prevailing syntactic theory will have a certain number of rules based on how many distinct (in whatever relevant sense) properties are found in describing the language. For example, 20 apparent rules might be combined into 10 with the same empirical coverage for a given theoretical perspective, presumably to a similar extent cross-linguistically. Thus we may ask not just *how many* but also *what types* of rules are found. But counting unique grammatical properties would require exhaustive descriptions for each language, so we must estimate the probability of a linguist identifying more distinct properties in one language than another.

2.2. Dependency Density

A preliminary proposal, appropriately convenient for the data provided for this task, would be to consider the types and distribution of syntactic dependencies in a corpus for each language. For example, adjectives may modify nouns, and subjects may indicate the agent of verbs. Given part-of-speech tagging along with dependency information in the corpus, we can measure the total number of unique dependency relationships. And by looking at the same amount of data for each language, we can estimate *dependency density*. Languages with higher dependency densities have more possible constructions for linguists to investigate, at least some of which may have unique properties that need to be explained independently, regardless of the particular theoretical framework. It is important to note that the most basic dependency types (adjective-noun, subject-verb, etc.) will be both frequent in a given language, and also most likely to be found in all of the languages in the sample. Therefore, we must try to identify infrequent, typologically unusual syntactic features not found in all languages. By considering each unique dependency relationship regardless of frequency, this metric of comparative complexity will be primarily determined by the more numerous *infrequent* constructions in the language, given that the more common constructions will be shared, balancing out across the languages.

The complexity measurement for each language was calculated from the tagged corpus data with a triplet for each word: the part-of-speech (UPOS); the dependency relation (DEPREL); and the part-of-speech of that related word. Lexical information was discarded, as well as punctuation. The first 36,000 dependencies of this type were considered in the corpus for each language, limited by the smallest corpus (Hungarian: 36,225 dependencies available).[3]

> **Results** (low to high complexity): Hindi (306); Slovenian (366); Bulgarian (374); Vietnamese (374); Polish (392); Italian (399); Urdu (406); Galician (411); Greek (419); Persian (448); Estonian (461); Norwegian (Bokmaal) (508); Norwegian (Nynorsk) (515); French (515); Portuguese (529); Danish (535); Swedish (537); Catalan (543); Slovak (544); Chinese (550); Serbian (564); Spanish (572); Afrikaans (576); Ukrainian (582); Russian (588); Arabic (598); Hungarian (606); Finnish (609); Czech (618); Basque (626); Latvian (643); Turkish (653); Hebrew (664); Romanian (703); Dutch (744); Croatian (749); English (763)

Thus, *based on this data alone*, a linguist writing a grammar would have more constructions to explain for English than Hindi, and presumably some of those constructions would require unique explanations. These results must be interpreted tentatively as we have no independent metric to test their validity.[4]

2.3. The Zipfian problem

The available corpora are of limited size, and the difference between languages might be based on frequency distribution of dependency relationships rather than whether particular dependencies exist at all in the language, a problem exaggerated by varied text types in the data (from long paragraphs to abbreviated internet comments). We would hope that the results would be replicable with more, and larger data sets, but this is uncertain. As Zipf (1935) found for the distribution of lexical items, the distribution of syntactic constructions is logarithmic and biased toward the most frequent items (Köhler, 2007).[5] Looking at the the largest data set (Czech: 1.29 million dependencies), new unique dependency relationships are progressively rarer, but there is no

[3] This translates to 3,280 sentences or 35,259 words for English, for example. The figures for the other languages vary, as there may be more or fewer dependencies per sentence in each language.

[4] Encouragingly, some of the closely related languages, such as the two varieties of Norwegian, are ranked similarly. Additionally, if we instead measure bigrams (part-of-speech pairs, based on adjacency in the text, still setting aside lexical information but now also dependency parsing), the results are statistically correlated with the dependency rankings (r^2=0.11; p<.05), although the ranking of individual languages varies. Including lexical information eliminates that correlation (but results for lexical dependencies and lexical bigrams also appear similar to each other). This suggests that a measurement of syntactic complexity (without influence of lexical density) would require at least a tagged corpus, but possibly not a dependency-parsed corpus.

[5] Another possible approach would be to consider *only* infrequent types, discarding information about the more frequent dependencies in the corpus data. However, because there are more infrequent than frequent constructions, the distribution are still broadly statistically similar.

indication that all of them have been found by the end of even this large corpus. Compared to the 618 unique dependency relationships among 36,000, there are 1,538 in the full corpus of 1.29 million. At the very least, we must conclude that much larger corpora are required for representative measurements.

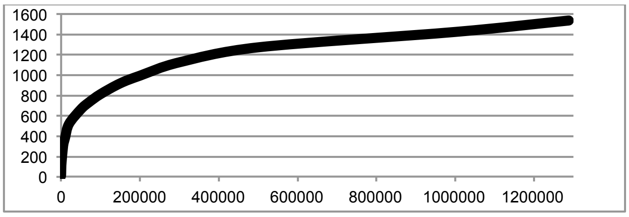

Figure 1. Cumulative unique dependency relationships per total number in corpus (Czech).

3. Typological considerations

The top-down approach presented above can be contrasted with a bottom-up approach based on linguistic analysis of the features of individual languages. In this sense, we can apply the specific construction type discussed by Ross (2014), namely *verbal pseudocoordination* (PC, such as English *go and get* or *try and do*), where a dependency relationship between two verbs is indicated by an anomalous use of the coordinating conjunction *and* (which importantly would not be tagged as such in a corpus). Later typological surveys (Ross 2016, 2017) provide the relevant data for this comparison. As it is especially common in Europe, PC is found in most languages of this biased sample. (The 7 without PC are: Chinese, Dutch, French, Hindi, Slovenian, Urdu and Vietnamese.) A linguist describing the 30 languages with PC would need to explain this feature and any idiosyncrasies it has (see the arguments in Ross, 2014); the number of different PC constructions in each language could also be considered, ranging from just one to many types. Additionally, just as some languages have recently developed PC, Dutch (Van Pottelberge, 2002) and Chinese (Tsai, 2007) had PC historically, an apparent loss of complexity. However, PC has been functionally replaced by other syntactic constructions, such as infinitives in Dutch and Serial Verb Constructions (SVCs) in Chinese. In fact, SVCs should be similarly considered because they represent complex *but unmarked* relationships between verbs (Escure, 2009), and some languages have both PC and SVCs. As they are not a typical feature in Europe, SVCs are rare in the current sample, with extensive usage only in Chinese and Vietnamese. (Following and expanding on Ross et al. 2015, the other languages with limited usage of SVCs in the sample are Afrikaans, Arabic, Estonian, Hindi, Hungarian, Persian, Russian, Turkish, and Urdu, and marginal usage in Basque and English). See Ross (forthcoming) for the distribution of SVCs, PC and related syntactic features.

To determine the overall complexity of languages, many more features should be considered. A full study of this sort would require extensive documentation and linguistic analysis for each language in a sample. However, we can attempt to estimate the distribution based on available typological databases, such as the *World Atlas of Language Structures* (WALS: Haspelmath et al. 2005). In fact, Bentz et al. (2016) found strong correlations between several automated metrics and features from WALS, though they only considered morphological complexity. Although WALS offers over 140 feature sets, only a small subset are relevant to measuring differential complexity cross-linguistically. Among the syntactic features in WALS, most (such as word order features) represent variation but not one language having more properties than another; therefore, only 8 relevant features were selected: gender (30A); articles (37A/38A); case (49A/51A); having two basic word orders (81A/81B); passives (107A/108A); syntactic expression of negation (112A); syntactic expression of polar question (116A); and copula omission (120A). These were coded as binary features (1=presence; 0=absence), and also including the additional data for PC and SVCs, estimated syntactic complexity was calculated as an average of these 10 features. The full results are presented in the accompanying materials. However, this was found to be a relatively weak measure of syntactic complexity for several reasons: (1) the limited number of variables available; (2) the similar distribution of many of these features in the (mostly European) languages in the sample; and gaps in the data for some languages in WALS (Afrikaans, Galician, and Slovak should be removed for lack of data). Furthermore, this sort of large-scale typological database lends itself to widespread features, rather than any unique properties of individual languages, thus obscuring complexity, given that infrequent or unusual features will account for the majority of a native speaker's knowledge, as discussed above. Therefore, specific annotation by experts of features in each language is desired.

Finally, let us consider the possibility of measuring complexity based literally on the length of published descriptive grammars (number of pages), as mentioned metaphorically above. A ranking based on the most detailed available grammar for each language is presented in the accompanying materials.

Unfortunately, but not surprisingly, no statistical correlation was found between any pair among the corpus-based dependency metric, the 10-feature WALS metric, or the page count of descriptive grammars. Whether there can be any correlation between bottom-up and top-down methods remains to be seen.

References

Bentz, C., Ruzsics, T., Koplenig, A., & Samardžić, T. (2016). A Comparison Between Morphological Complexity Measures: Typological Data vs. Language Corpora. In D. Brunato, F. Dell'Orletta, G. Venturi, T. François, & P. Blache (Eds.), *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity* (pp. 142–153). Osaka: COLING 2016 Organizing Committee. http://aclweb.org/anthology/W16-41

Chomsky, N. (1995). *The Minimalist program*. Cambridge, MA: MIT Press.

Escure, G. (2009). Is verb serialization simple? Evidence from Chinese Pidgin English. In N. Faraclas & T. B. Klein (Eds.), *Simplicity and Complexity in Creoles and Pidgins*. London: Battlebridge.

Gell-Mann, M. (1994). *The quark and the jaguar: adventures in the simple and the complex*. New York: W.H. Freeman & Co.

Goldberg, A. E. (1995). *Constructions: a construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Haspelmath, M., Dryer, M. S., Gil, D., & Comrie, B. (Eds.). (2005). *World Atlas of Language Structures*. Oxford: Oxford University Press. http://wals.info/

Köhler, R. (2007). Quantitative Analysis of Syntactic Structures in the Framework of Synergetic Linguistics. In A. Mehler & R. Köhler (Eds.), *Aspects of Automatic Text Analysis* (pp. 191–209). Berlin: Springer.

Ross, D. (2014). The importance of exhaustive description in measuring linguistic complexity: The case of English *try and* pseudocoordination. In F. J. Newmeyer & L. B. Preston (Eds.), *Measuring Grammatical Complexity* (pp. 202–216). Oxford: Oxford University Press.

Ross, D. (2016). Between coordination and subordination: Typological, structural and diachronic perspectives on pseudocoordination. In F. Pratas, S. Pereira, & C. Pinto (Eds.), *Coordination and Subordination: Form and Meaning — Selected Papers from CSI Lisbon 2014* (pp. 209–243). Newcastle upon Tyne: Cambridge Scholars Publishing.

Ross, D. (2017). Pseudocoordinación del tipo tomar y en Eurasia: 50 años después [Pseudocoordination with take and in Eurasia: 50 years later]. Presented at Lingüística Coseriana VI, Lima, Peru.

Ross, D. (forthcoming). *Pseudocoordination, serial verb constructions and multi-verb predicates: The relationship between form and structure* (Ph.D. dissertation). University of Illinois at Urbana-Champaign, Urbana, IL.

Tsai, W.-T. D. (2007). Conjunctive Reduction and its Origin: A Comparative Study of Tsou, Amis, and Squliq Atayal. *Oceanic Linguistics*, *46*(2), 585–602.

Van Pottelberge, J. (2002). Nederlandse progressiefconstructies met werkwoorden van lichaamshouding: specificiteit en geschiedenis. *Nederlandse Taalkunde*, *7*(2), 142–174.

Zipf, G. K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. Boston: Houghton Mifflin.