Details matter: problems and possibilities for measuring cross-linguistic complexity

Daniel Ross

djross3@gmail.com — danielrosslinguist.com

University of Illinois at Urbana-Champaign

Shared task workshop on

Measuring Language Complexity (MLC2018)

April 15th, 2018, Torun, Poland

Goals

- Measure the linguistic complexity of 37 languages for this shared task workshop
  - Corpora from Universal Dependencies project provided for:
    Afrikaans, Arabic, Basque, Bulgarian, Catalan, Chinese, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, Galician, Greek, Hebrew, Hindi, Hungarian, Italian, Latvian, Norwegian-Bokmaal, Norwegian-Nynorsk, Persian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish, Turkish, Ukrainian, Urdu, Vietnamese
- Evaluate the results from a typological perspective
- Consider implications for methodology and linguistic theory

How to measure linguistic complexity?

- No standard definition for linguistic complexity (Miestamo, Sinnemäki & Karlsson 2008; Sampson, Gil & Trudgill 2009; Newmeyer & Preston 2014; Baechler & Seiler 2016)
- Therefore no standard metric to evaluate results as correct/insightful or not
- What to do with corpus data and results?
  - *This shared task is a good first step!*

Approaches to complexity

- Grammatical complexity: knowledge of speakers
- Usage complexity: production of speakers
- Psycholinguistic complexity: processing effort
- Acquisition complexity: difficulty for learners (L1/L2)

- Also: overall/average complexity vs. identification of 'most complex' feature(s)

Ross 2014: Defining grammatical complexity

- Consider all details of a grammatical system, independent of frequency of use
- Measure *effective complexity* (Gell-Mann 1994:58)
  - Similar to Kolmogorov complexity, but with stochastic information (i.e., vocabulary) removed; patterns remain
  - Ross (2014) argued for this *theoretically*; now let's test it!
- Thought experiment: grammatical complexity should be equivalent to description length for an optimally written descriptive grammar

Literal thought experiment

- As a **baseline**, *how long are* descriptive grammars?
  - Most detailed available grammars selected based on author's own experience and Glottolog database
  - Length measured as simple page count
    - See Supplementary Materials for details and discussion

Urdu (300), Ukrainian (315), Hindi (318), Serbian (321), Czech (338), Croatian (362), Persian (379), Bulgarian (409), Estonian (413), Dutch (438), Hungarian (472), Vietnamese (478), Slovenian (494), Turkish (575), Hebrew (580), Slovak (583), Polish (647), Greek (648), Afrikaans (652), Portuguese (718), French (786), Arabic (812), Chinese (847), Romanian (851), Basque (943), Latvian (1024), Norwegian-Bokmaal (1223), Norwegian-Nynorsk (1223), Russian (1421), Catalan (1439), Galician (1641), Finnish (1698), Danish (1842), English (1842), Italian (2351), Swedish (2656), Spanish (4417)

Measuring grammatical complexity

- Grammatical complexity is independent of frequency of usage: all grammatical patterns known to native speakers contribute to overall measure
- Grammatical complexity should be predicted by the **number of rules** in a grammatical description
  - The rules themselves may be more or less complex, but the first step is at least identifying all the relevant rules
- Grammatical theories vary substantially in types of rules, so to begin we should identify all relevant data

Some comments on other workshop papers

- No standard definition of linguistic complexity
  - We cannot evaluate other metrics as right or wrong
  - We can look for correlations
  - We can discuss methodologies
- The other papers do not in general attempt to describe *grammatical complexity*
  - They instead are dependent on frequency of usage
  - The other papers investigate *usage complexity*
  - Indeed more relevant and easier for a corpus-based task

Grammatical complexity measured in corpora?

- Can corpora help for grammatical complexity?
- On the surface, corpora show *usage*, not *grammar*
- We must measure how many rules are required to explain the data, not how often those rules are used
- Can we *estimate* grammatical complexity based on the number of distinct rules found in corpus data?
  - Theories vary in how to interpret those rules, but all distinct properties of the grammar must be explained

Dependency density

- The data provided for this shared task is ideal for comparing dependencies cross-linguistically
  - *I consider only syntactic complexity in this study*
- Many dependency types would be found in most or all languages (subject-verb, adjective-noun)
- But languages with more dependency types would presumably be those with more syntactic rules!
  - Some rules might explain multiple dependency types, but number of dependencies should correlate with number of rules

Measuring dependencies

- Dependencies were counted as triplets:
  - The part-of-speech (UPOS) for each word
  - The dependency relation (DEPREL) to another word
  - The part-of-speech of that related word
    - Lexical information was discarded (as well as punctuation)
- First 36,000 dependencies for each language
  - Limited by smallest corpus (Hungarian: 36,225 dependencies available)
  - For example, equivalent to 3,280 sentences or 35,259 words for English corpus; varies
- Number of unique dependencies as complexity

Dependency results

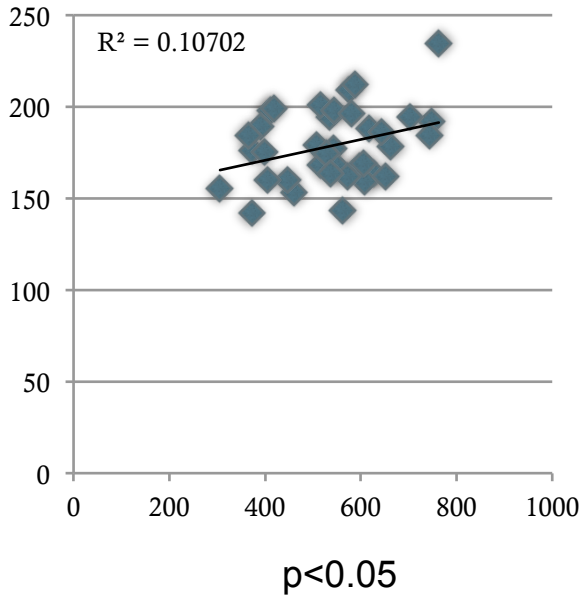
- Unique dependencies found for each language:
  Hindi (306); Slovenian (366); Bulgarian (374); Vietnamese (374); Polish (392); Italian (399); Urdu (406); Galician (411); Greek (419); Persian (448); Estonian (461); Norwegian (Bokmaal) (508); Norwegian (Nynorsk) (515); French (515); Portuguese (529); 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)
- These results suggest that a descriptive linguist:
  - Would write the shortest grammar of Hindi
  - And the longest grammar of English

Discussion of results

- A wide range of results suggests a strong measure
  - Lowest: 306 dependencies; highest: 763
- Similar results for some closely related languages
  - Norwegian varieties; Russian & Ukrainian; Swedish & Danish; etc.
- Difficult to know if overall rankings are relevant without a baseline for comparison
  - Why are some families split?
    - Hindi & Urdu? Italian & Spanish? Norwegian & Swedish?

Correlation with bigrams

- When looking only at part-of-speech, there is a statistical correlation between number of unique dependency triplets and number of unique part-of-speech bigrams (based on linear order only)
- Could a tagged corpus substitute for a parsed corpus?
  - Correlation is significant, but not precise



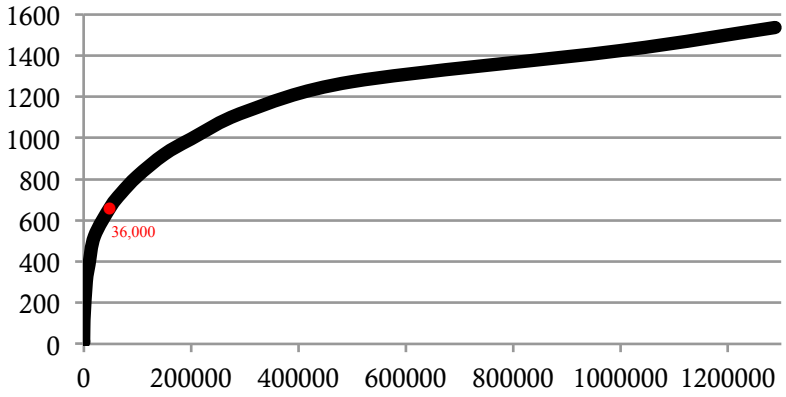
Workshop comparisons: also strong statistical correlation ($p<0.005$) of dependencies with CR_POSP measure for example.

The Zipfian problem

- Corpora limited in size, and cross-linguistic differences could be due to frequency, not grammar
  - Would larger data sets give the same results?
- Zipf (1935) found lexical items are distributed logarithmically: the most frequent types account for substantially more of the tokens in a corpus
- The same pattern is found for syntactic structures (Köhler 2007); using small corpora is a problem

Dependencies in larger corpora

- The Czech corpus is substantially larger than the others, with 1.29 million dependencies total
  - versus 36,000 considered for each language above (red dot)
  - As the corpus grows, unique dependencies continue to be found, though less frequently
  - No indication even this corpus is large enough to find all dependencies



Typological considerations

- How does bottom-up typological research assessing the distribution of grammatical features compare to the top-down corpus-based approach?
- Let us consider the specific example from Ross (2014) of verbal pseudocoordination (PC)
  - Ex: English *go and get* or *try and do*
  - A dependency relationship expressed via anomalous usage of the coordinating conjunction *and*
    - *Note that a corpus would not have* and *tagged in such a way*

Pseudocoordination (PC)

- PC in some languages is associated with unusual morphosyntactic properties (Ross 2014 on *try and*)
  - Therefore, the presence of PC in a language may indicate a longer description length: a topic to discuss in a grammar
    - The number of distinct subtypes of PC should also be considered
  - Only one of many features to consider, but not to be ignored
- PC rare but common in Europe (Ross 2016, 2017), our current sample is European-biased:
  - 7 languages in the sample lack PC: Chinese, Dutch, French, Hindi, Slovenian, Urdu and Vietnamese

PC and other constructions

- Looking at individual features allows for a diachronic perspective on complexity:
- Some languages historically had PC but no longer do
  - Ex: Dutch (Van Pottelberge 2002); Chinese (Tsai 2007)
- Do these languages exemplify *decreasing complexity?*
  - Not necessarily: PC has been functionally replaced
    - Dutch: Infinitives; Chinese: Serial Verb Constructions (SVCs)
  - Thus, more features beyond PC to consider…

Serial Verb Constructions

- SVCs are also important for measuring complexity
  - complex *but unmarked* relationships between verbs (Escure 2009)
    - Similar to the English construction *go get* (like PC, but without *and*)
- Some languages have both PC and SVCs
  - SVCs rare in Europe and the current sample
    - Extensive usage only in Chinese and Vietnamese
    - Following the definition in Ross et al. (2015), SVCs found also in Afrikaans, Arabic, Estonian, Hindi, Hungarian, Persian, Russian, Turkish and Urdu; marginal usage in English and Basque
  - *See Ross (forthcoming) on distribution of PC, SVCs and related features*

WALS: Syntactic typology

- PC and SVCs are just two of many features to be considered for a typological measure of complexity
- The *World Atlas of Language Structures* (WALS: Haspelmath et al. 2005) is a good starting point
  - See Bentz et al. (2016) on morphological complexity
  - WALS has over 140 features, but few relevant here:
    - Only a subset are related to syntax in particular
    - Most syntactic features are *variation* (like word order), not presence/absence of features showing different complexity

WALS metric

- 8 features in WALS were found to relate to differential syntactic complexity *(see paper for details)*
  - WALS also biased toward common features cross-linguistically
- Combined with 2 features of PC and SVCs above
- Each feature coded as binary value, averaged:
  - 1 if feature found (and contributing to complexity)
  - 0 if not found
    - Ex: languages with definite and/or indefinite articles were considered more complex than those without

WALS results

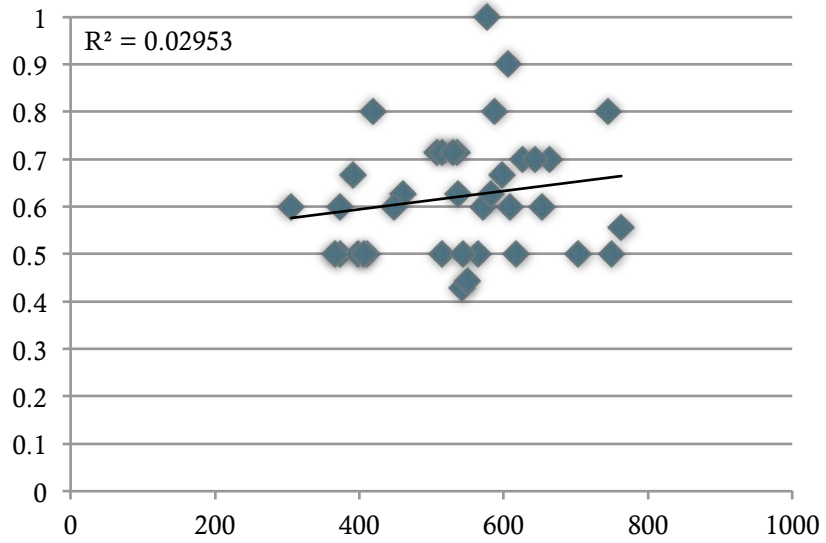
- Average of the 10 features gives a preliminary measure of typological syntactic complexity:

Catalan (0.429), Chinese (0.444), Bulgarian (0.5), Croatian (0.5), Czech (0.5), French (0.5), Galician (0.5), Italian (0.5), Romanian (0.5), Serbian (0.5), Slovak (0.5), Slovenian (0.5), Urdu (0.5), English (0.556), Finnish (0.6), Hindi (0.6), Persian (0.6), Spanish (0.6), Turkish (0.6), Vietnamese (0.6), Estonian (0.625), Swedish (0.625), Ukrainian (0.625), Arabic (0.667), Polish (0.667), Basque (0.7), Hebrew (0.7), Latvian (0.7), Danish (0.714), Norwegian-Bokmaal (0.714), Norwegian-Nynorsk (0.714), Portuguese (0.714), Dutch (0.8), Greek (0.8), Russian (0.8), Hungarian (0.9), Afrikaans (1)

  - Some values missing for some languages in WALS
- More data is needed for a representative comparison

Comparison of metrics

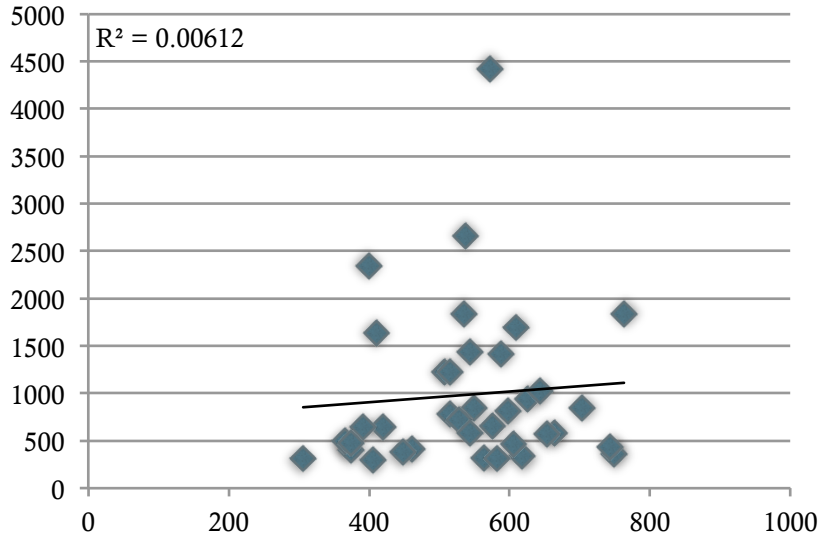
- Dependencies vs. WALS



  - No statistical correlation

Comparison of metrics

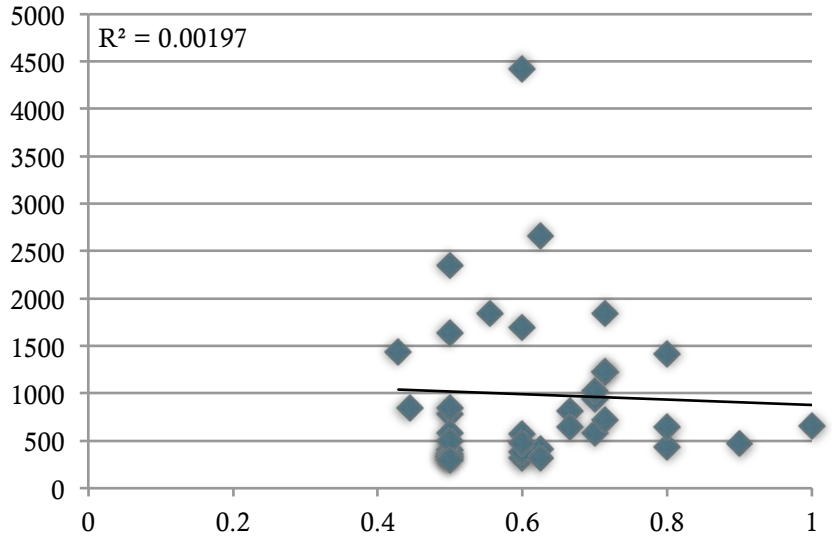
- Dependencies vs. Descriptive Grammar Length



  - No statistical correlation

Comparison of metrics

- WALS vs. Descriptive Grammar Length



  - No statistical correlation

Comparison of metrics

- No statistical correlation between the three metrics
  - Not surprising given the limited data in each
    - Unclear whether there should be a top-down/bottom-up correlation for these metrics
- Each metric is insufficient:
  - The corpus size was too small to be reliable for approximating the distribution of infrequent features
    - But would larger corpora represent more typos in unique configurations?
  - Too few typological variables to estimate syntactic complexity
  - Current descriptive grammars are far from exhaustive descriptions

Conclusions

- Corpus-based measures of linguistic complexity must be based on much larger corpora
  - Additionally, for comparability we must distinguish between *grammatical* and *usage* complexity measures
  - A more diverse sample of languages could also be helpful
- What is the best metric proposed here?
  - Probably the length of descriptive grammars, because it represents the state of the art of our knowledge about languages
  - As argued in Ross (2014), exhaustive description is crucial
- No reason to believe all languages have equal complexity
  - But we cannot yet identify the potentially small differences

References

Baechler, Raffaela & Guido Seiler (eds.). 2016. *Complexity, isolation, and variation*. Berlin: De Gruyter.

Bentz, Christian, Tatyana Ruzsics, Alexander Koplenig & Tanja Samardžić. 2016. A Comparison Between Morphological Complexity Measures: Typological Data vs. Language Corpora. In Brunato et al. (eds.), *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, 142–153. Osaka, Japan: COLING 2016 Organizing Committee. http://aclweb.org/anthology/W16-41.

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

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

Haspelmath, Martin, Matthew S. Dryer, David Gil & Bernard Comrie (eds.). 2005. *World Atlas of Language Structures*. Oxford: Oxford University Press. http://wals.info/.

Miestamo, Matti, Kaius Sinnemäki & Fred Karlsson (eds.). 2008. *Language complexity: typology, contact, change*. Amsterdam: John Benjamins.

Newmeyer, Frederick J. & Laurel B. Preston (eds.). 2014. *Measuring Grammatical Complexity*. Oxford: Oxford University Press.

Ross, Daniel. 2014. The importance of exhaustive description in measuring linguistic complexity: The case of English *try and* pseudocoordination. In Newmeyer & Preston (eds.), 202–216.

References

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

Ross, Daniel. 2017. Pseudocoordinación del tipo tomar y en Eurasia: 50 años después. Presented at *Lingüística Coseriana VI*, Lima, Peru.

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

Ross, Daniel, Ryan Grunow, Kelsey Lac, George Jabbour & Jack Dempsey. 2015. Serial Verb Constructions: a distributional and typological perspective. Presented at *Illinois Language and Linguistics Society (ILLS) 7*, Urbana, Illinois. http://hdl.handle.net/2142/88844.

Sampson, Geoffrey, David Gil & Peter Trudgill (eds.). 2009. *Language complexity as an evolving variable*. New York: Oxford University Press.

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

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