KOLMOGOROV COMPLEXITY AS A UNIVERSAL MEASURE OF LANGUAGE COMPLEXITY

Katharina Ehret

kehret@sfu.ca

Department of Linguistics, Simon Fraser University, Burnaby, Canada

This paper presents an unsupervised information-theoretic measure that is a promising candidate for becoming a universally applicable metric of language complexity. The measure boils down to Kolmogorov complexity and uses compression programs to assess the complexity in text samples via their information content. Generally, better compression rates indicate lower complexity. In this paper, the measure is applied to a typological dataset of 37 languages covering 7 different language families. Specifically, overall, morphological and syntactic complexity are measured. The results often coincide with intuitive complexity judgements, e.g. Afrikaans is overall comparatively simple, Turkish is morphologically complex. Yet, in some cases the results are surprising, e.g. Chinese turns out to be morphologically highly complex. It is concluded that the method needs further adaptation for the application to different writing systems. Despite this caveat, the method is in principle applicable to all types of languages.

1. Introduction

Language complexity is a very fashionable research topic in the typological-sociolinguistics community (Baechler & Seiler, 2016; Baerman, Brown, & Corbett, 2015; Kortmann & Szmrecsanyi, 2012; Sampson, 2009; Miestamo, 2008). Theoretical complexity research is concerned with the definition and measurement of language complexity, and the reasons for variation in language complexity. Most of this research analyses complexity variation in cross-linguistic datasets (e.g. Nichols, 1992) or different varieties of the same language (e.g. Szmrecsanyi, 2009; Trudgill, 2009). Despite the plethora of research on language complexity, no universally applicable definition or metric of complexity exists. Thus, it is virtually impossible to compare complexity measurements across different studies.

Against this backdrop, this paper presents an unsupervised information-theoretic measure of language complexity, which has the potential of becoming a universally applicable metric of complexity. This measure, also dubbed the compression technique (see Ehret, 2017), was first introduced by Juola (1998) and substantially extended by Ehret (2017), Ehret and Szmrecsanyi (2016), and Ehret (2014). The measure is based on the notion of Kolmogorov complexity and measures the information content of a string by the length of the shortest possi-

ble description that is required to (re)construct the exact string (Li, Chen, Li, Ma, & Vitányi, 2004; Juola, 2008). The two strings below, for example, both count ten symbols. String (1-a) can be compressed to four symbols. In contrast, the shortest description of string (1-b) is the string itself, which counts ten symbols. String (1-a) is therefore less complex than string (1-b).

(1) a. pkpkpkpkpk (10 symbols) → 5×gh (4 symbols)
b. c4pk?9agy7 (10 symbols) → c4pk?9agy7 (10 symbols)

Although Kolmogorov complexity is uncomputable it can be conveniently approximated with text compression programs. The basic idea behind the compression technique is that text samples which can be compressed comparatively better are linguistically comparatively less complex. In linguistic terms, information-theoretic Kolmogorov-based complexity is a measure of structural surface redundancy and (ir)regularity. In contrast to most traditional complexity metrics which are often based on subjective or reductionist feature selection, the measure is arguably more objective and holistic, and at the same time inherently usage-based as it is radically text-based. In fact, it is agnostic about form-function pairings as the algorithm has no knowledge of the texts it is applied to. It is this text-based (in contrast to feature-based) approach that makes the compression technique a promising candidate for a universally applicable measure of language complexity. In this paper, the compression technique is used to measure overall and, through the application of various distortion techniques, morphological and syntactic complexity.

2. Methodology and data

The dataset is drawn from the Universal Dependencies project (v2.1) and specifically comprises a convenient sample of 37 languages covering 7 different language families: Afrikaans, Arabic, Basque, Bulgarian, Catalan, Chinese, Croation, Czech, Danish, Dutch, English, Estonian, Finnish, French, Galician, Greek, Hebrew, Hindi, Hungarian, Italian, Latvian, Norwegian Bokmaal, Norwegian Nyorsk, Persian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish, Turkish, Ukrainian, Urdu, Vietnamese. The current dataset thus consists of 37 text samples, one for each language. All texts were UNICODE normalised and converted to lowercase; non-alphabetical characters were automatically removed and all end-of-sentence markers were replaced by a single fullstop (for details see Ehret, 2017).

Overall complexity is measured in a straighforward manner by taking two measurements for each text sample: the file size (in bytes) before compression and the file size (in bytes) after compression. The file size pairings are then subjected to regression analysis in order to eliminate any trivial correlations between the two measurements. The resulting *adjusted overall complexity scores* (regression residuals, in bytes) are taken as indicator of the overall complexity of the text

samples. Higher scores indicate overall higher linguistic complexity; lower scores indicate lower complexity.

Inspired by Juola (1998, 2008), morphological and syntactic complexity are measured by applying distortion techniques prior to compression. Syntactic distortion is achieved by the deletion of 10% of all tokens in each text file. This disrupts word order regularities and greatly affects syntactically complex texts, i.e. texts with a comparatively fixed word order. Syntactically less complex texts are little affected by this procedure, as they lack syntactic interdependencies that could be compromised. Comparatively bad compression ratios after syntactic distortion indicate comparatively high syntactic complexity. Morphological distortion is performed by the deletion of 10% of all characters in each text file thereby creating new "word forms". This compromises morphological regularity: morphologically complex languages exhibit overall a relatively large amount of word forms in any case, so they are little affected. Yet, in morphologically less complex languages proportionally more random noise is created. Comparatively bad compression ratios after morphological distortion thus indicate low morphological complexity. In this spirit, the scores for morphological and syntactic complexity are calculated based on two file sizes: the compressed file size of the original text and the compressed file size of the distorted text. To be specific, the *morphological complexity score* is defined as $-\frac{m}{c}$, where m is the compressed file size after morphological distortion and c the original compressed file size. The *syntactic complexity score* is defined as $\frac{s}{c}$, where s is the compressed file size after syntactic distortion and c the file size before distortion.

The above described distortion and compression procedure uses gzip (v1.2.4 `http://www.gzip.org/`) for text compression, and is applied with $N = 1000$ iterations (for details see Ehret, 2017).[1] All complexity scores reported in this paper are based on the arithmetic mean calculated for the individual complexity scores across $N = 1000$ iterations. Detailed statistics such as individual complexity scores and file sizes are included in the supplementary material. All statistics were conducted in R (v3.3.3, R Core Team (2017)).

3. Kolmogorov complexity in a typological perspective

In Fig. 1 (upper plot) an overall complexity hierarchy of the 37 languages is presented. In many cases, the results match with general expectations about complexity. For example, the Afrikaans text is overall less complex than the Hungarian text; the English text is overall below-average complex, while the French text is overall above-average complex. In some cases, however, the compression results are surprising: Chinese, in particular, is an outlier in the dataset. Its ranking as the overall most complex text is most likely an artifact of its specific writing sys-

[1] The compression and distortion scripts are available at `https://github.com/katehret/measuring-language-complexity`.

tem. In a similar vein, Urdu is ranked as one of the overall most complex texts, while Hindi is ranked as the overall least complex text. The placement of Urdu and Hindi at the extreme opposite ends of the overall complexity hierarchy could also be due to their use of different writing systems.

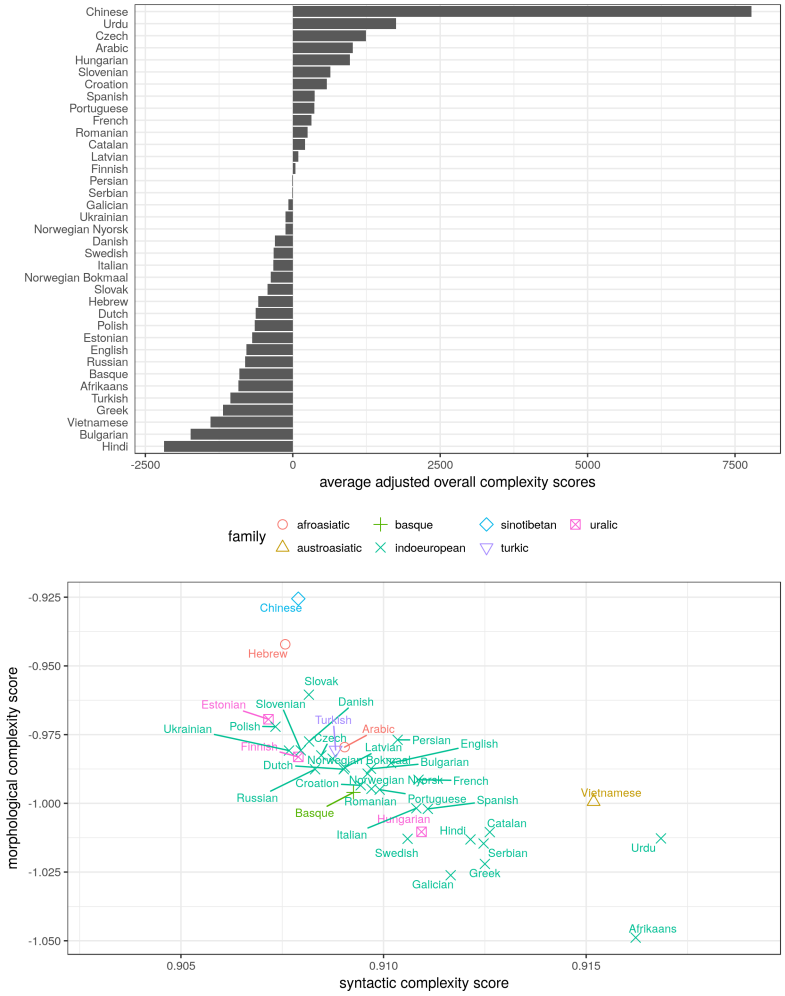

Figure 1. Upper plot: Overall complexity hierarchy. Negative residuals indicate below-average complexity; positive residuals indicate above-average complexity. Lower plot: Morphological by syntactic complexity. Abscissa indexes increased syntactic complexity; ordinate indexes increased morphological complexity.

The lower plot of Fig. 1, displays the compression measurements in the two-dimensional space of morphological and syntactic Kolmogorov complexity. Generally, the results coincide with intuitive complexity judgements. The Afrikaans text, for instance, exhibits the least morphological complexity, i.e. it contains little word form variation. In terms of syntax the Afrikaans text is rather complex, i.e. it has lots of word order rules and comparatively rigid syntactic patterns. The Hebrew text, in contrast, is comparatively more complex in terms of morphology and exhibits average syntactic complexity. Yet, some complexity placements are rather counter-intuitive: For example, the English text is morphologically more complex than the Hungarian text. This dislocation must be attributed to a lack of content control in the data as the compression technique has been shown to reliably measure complexity in typological datasets (Ehret & Szmrecsanyi, 2016). Chinese, again, is an outlier in the dataset, and exhibits the highest morphological complexity.

4. Conclusion

This paper presents Kolmogorov complexity as a universal measure of language complexity which could facilitate the comparison of complexity measurements across different studies. That said, in its current implementation the compression technique relies on distortion procedures developed for the Latin alphabet; this operationalisation is problematic for languages like Chinese. Future applications should utilise more universally applicable distortion techniques (see e.g. Koplenig, Meyer, Wolfer, & Müller-Spitzer, 2017). Furthermore, the comparability and reliability of the results obtained by the compression technique greatly depend on the quality of the input. Specifically, the comparability of the propositional content across different text samples is a major factor influencing the compression results (for a discussion see Ehret, 2017). For the analysis of large-scale typological datasets it is recommended to draw on parallel text corpora, such as the Bible, because differences due to propositional content can be ruled out (Wälchli, 2007), or on carefully compiled naturalistic datasets. Nevertheless, the compression technique is a promising candidate for becoming a universally applicable measure of language complexity because it does not rely on language-specific feature catalogues but is, in principle, applicable to all types of languages.

Acknowledgements

I am grateful to the Cusanuswerk (Bonn, Germany) for a generous PhD scholarship, and the Alexander von Humboldt Foundation (Bonn, Germany) for postdoctoral funding through a Feodor-Lynen Fellowship. My thanks go to Alexander Koplenig for help with UNICODE normalisation, and to Aleksandrs Berdicevskis and Christian Bentz for helpful comments and feedback. The usual disclaimers apply.

References

Baechler, R., & Seiler, G. (Eds.). (2016). *Complexity, Isolation, and Variation.* Berlin, Boston: De Gruyter.

Baerman, M., Brown, D., & Corbett, G. G. (Eds.). (2015). *Understanding and measuring morphological complexity.* New York: Oxford University Press.

Ehret, K. (2014). Kolmogorov complexity of morphs and constructions in English. *Language Issues in Linguistic Technology*, *11*, 43–71.

Ehret, K. (2017). *An information-theoretic approach to language complexity: variation in naturalistic corpora.* PhD dissertation, Freiburg.

Ehret, K., & Szmrecsanyi, B. (2016). An information-theoretic approach to assess linguistic complexity. In R. Baechler & G. Seiler (Eds.), *Complexity and isolation* (pp. 71–94). Berlin: de Gruyter.

Juola, P. (1998). Measuring linguistic complexity: the morphological tier. *Journal of Quantitative Linguistics*, *5*(3), 206–213.

Juola, P. (2008). Assessing linguistic complexity. In M. Miestamo, K. Sinnemäki, & F. Karlsson (Eds.), *Language Complexity: Typology, Contact, Change* (pp. 89–107). Amsterdam, Philadelphia: Benjamins.

Koplenig, A., Meyer, P., Wolfer, S., & Müller-Spitzer, C. (2017). The statistical trade-off between word order and word structure a - Large-scale evidence for the principle of least effort. *PLOS ONE*, *12*(3), e0173614.

Kortmann, B., & Szmrecsanyi, B. (Eds.). (2012). *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact.* Berlin/Boston: Walter de Gruyter.

Li, M., Chen, X., Li, X., Ma, B., & Vitányi, P. M. B. (2004). The similarity metric. *IEEE Transactions on Information Theory*, *50*(12), 3250–3264.

Miestamo, M. (2008). Grammatical complexity in a cross-linguistic perspective. In M. Miestamo, K. Sinnemäki, & F. Karlsson (Eds.), *Language Complexity: Typology, Contact, Change* (pp. 23–41). Amsterdam, Philadelphia: Benjamins.

Nichols, J. (1992). *Linguistic Diversity in Space and Time.* Chicago: University of Chicago Press.

R Core Team. (2017). *R: A language and environment for statistical computing.* Vienna, Austria.

Sampson, G. (2009). A linguistic axiom challenged. In G. Sampson, D. Gil, & P. Trudgill (Eds.), *Language Complexity as an Evolving Variable* (pp. 1–18). Oxford: Oxford University Press.

Szmrecsanyi, B. (2009). Typological parameters of intralingual variability: Grammatical analyticity versus syntheticity in varieties of English. *Language Variation and Change*, *21*(3), 319–353.

Trudgill, P. (2009). Vernacular Universals and the Sociolinguistic Typology of English dialects. In M. Filppula, J. Klemola, & H. Paulasto (Eds.), *Vernac-*

ular universals and language contacts : evidence from varieties of English and beyond (pp. 304–322). New York: Routledge.

Wälchli, B. (2007). Advantages and disadvantages of using parallel texts in typological investigations. *Language Typology and Universals*, *60*(2), 118–134.