Proceedings of the First Shared Task on Measuring Language Complexity

April 15, 2018
Toruń, Poland

Editors

Aleksandrs Berdicevskis
Christian Bentz

ISBN: 978-91-639-7435-9

Sponsors


EVOLANG


Kungl.
KUNGL. VITTERHETS HISTORIE
OCH ANTIKVITETS AKADEMIEN
Vitterhetsakademien


EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN


UPPSALA
UNIVERSITET


EVOLAEMP
LANGUAGE EVOLUTION: THE EMPIRICAL TURN


WORDS BONES GENES TOOLS
Tracking Linguistic, Cultural, and Biological Trajectories of the Human Past

PREFACE

An influential line of thinking within evolutionary linguistics is that languages change in response to socioecological pressures, i.e. adapt to their environmental niches. Language complexity is a common parameter to test for such adaptation. It is, however, notoriously difficult to define and measure. Virtually every study of complexity uses its own operationalization and measure. This can be problematic if measures yield different conclusions, since there currently is little consensus about how measures themselves can be evaluated and compared.

To overcome this, we organized this shared task on linguistic complexity. Shared tasks are widely used in computational linguistics, but this workshop, to our knowledge, is the first shared task in language typology and language evolution. The task was: measure and compare the complexities of a set of 37 language varieties of 7 families (see Table 1 and Figure 1). The participants were free to choose what they wanted to measure, but they were requested to clearly state: 1) what exactly is being measured; 2) how the measure is calculated, and the theoretical rationale behind the method; 3) the resulting value for each language. All corpus-based measures had to use the corpora available via the Universal Dependencies (UD) project, v2.1 (Nivre et al., 2017). There were no requirements about which level of annotation (if any) had to be used.

The workshop, hosted as a satellite event at the Evolang 12 conference, saw seven submissions that yielded 34 measures, addressing various facets of complexity and spanning phonetics, morphology, morphosyntax, syntax and semantics. Most of the measures are corpus-based. See the workshop's website (http://www.christianbentz.de/MLC_proceedings.html) for supplementary materials and values of all measures.

We would like to thank our sponsors, David Gil and Adam Schembri who kindly agreed to give invited talks at the workshop, and the developers of the UD corpora, in particular Joakim Nivre and Dan Zeman, who provided good advice on choosing suitable treebanks from the UD collection.

Aleksandrs Berdicevskis, Christian Bentz

References

Nivre, J., Željko, A., Ahrenberg, L. et al. (2017). Universal Dependencies 2.1, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-2515.

Table 1. Languages and treebanks used in the shared task.

Language	ISO	Treebank size (K)	Family	Genus
Afrikaans	afr	49	IE	Germanic
Arabic	arb	202	Afro-Asiatic	Semitic
Basque	eus	121	Basque	Basque
Bulgarian	bul	156	IE	Slavic
Catalan	cat	531	IE	Romance
Chinese	cmn	123	Sino-Tibetan	Sinitic
Croatian	hrv	197	IE	Slavic
Czech	ces	2222	IE	Slavic
Danish	dan	100	IE	Germanic
Dutch	nld	208	IE	Germanic
English	eng	254	IE	Germanic
Estonian	est	106	Uralic	Finnic
Finnish	fin	202	Uralic	Finnic
French	fra	402	IE	Romance
Galician	glg	138	IE	Romance
Greek	ell	63	IE	Greek
Hebrew	heb	161	Afro-Asiatic	Semitic
Hindi	hin	351	IE	Indic
Hungarian	hun	42	Uralic	Ugric
Italian	ita	293	IE	Romance
Latvian	lav	90	IE	Baltic
Norwegian (Bokmaal)	nob	310	IE	Germanic
Norwegian (Nynorsk)	nno	301	IE	Germanic
Persian	pes	152	IE	Iranian
Polish	pol	83	IE	Slavic
Portuguese	por	227	IE	Romance
Romanian	ron	218	IE	Romance
Russian	rus	1107	IE	Slavic
Serbian	srp	86	IE	Slavic
Slovak	slk	106	IE	Slavic
Slovenian	slv	140	IE	Slavic
Spanish	spa	549	IE	Romance
Swedish	swe	96	IE	Germanic
Turkish	tur	58	Turkic	Turkic
Ukrainian	ukr	100	IE	Slavic
Urdu	urd	138	IE	Indic
Vietnamese	vie	43	Austro-Asiatic	Viet-Muong

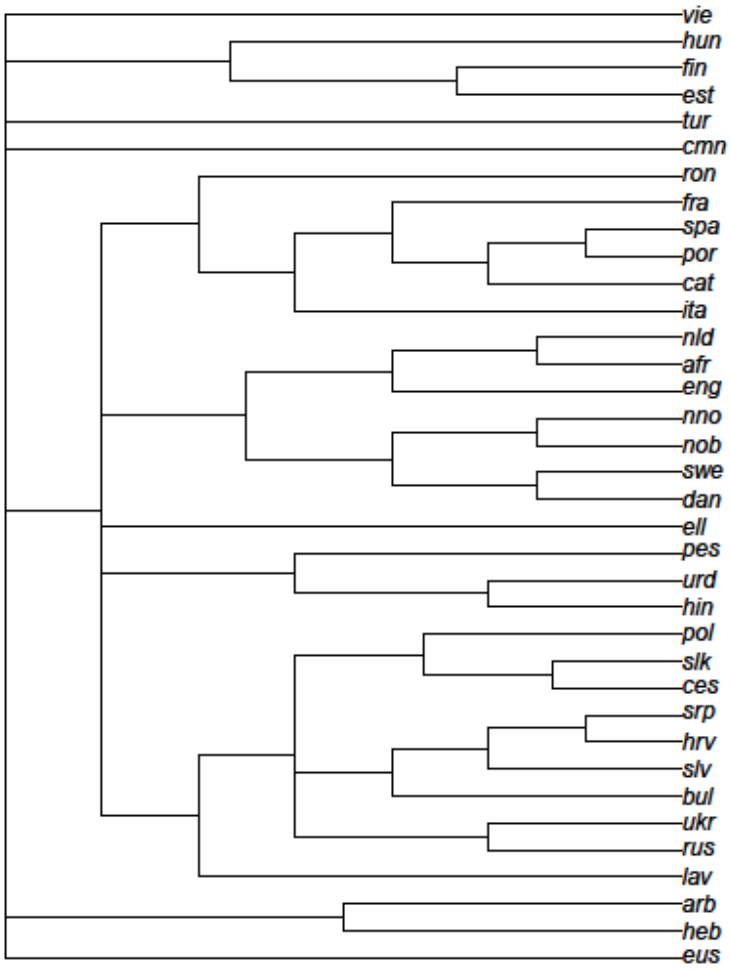

Figure 1. A phylogenetic tree of the 37 languages used in the shared task, based on the Glottolog classification. For illustration purposes the 7 families are rooted in the "World" node.

CONTENTS

EXPLOITING UNIVERSAL DEPENDENCIES TREEBANKS FOR MEASURING MORPHOSYNTACTIC COMPLEXITY

Çağrı Çöltekin[*1] and Taraka Rama[2]

[*]Corresponding Author: ccoltekin@sfs.uni-tuebingen.de
[1]Department of Linguistics, University of Tübingen, Germany
[2]Department of Informatics, University Oslo, Norway

We present six different measures of morphosyntactic complexity, calculated on 37 Universal Dependencies treebanks. We define the measures (some of which are not published in the earlier literature), present the results, and discuss relationships between the measures.

1. Introduction

There has been recent interest in quantifying linguistic complexity (Juola, 1998; Dahl, 2004; Newmeyer & Preston, 2014; Bentz, Alikaniotis, Cysouw, & Ferrer-i Cancho, 2017; Koplenig, Meyer, Wolfer, & Mueller-Spitzer, 2017; Stump, 2017). Besides the theoretical interest, quantifying complexity of languages or subsystems of languages is also important for first and second language acquisition research. In this paper, we present a number of morphosyntactic measures, some proposed in earlier literature, and some novel to the best of our knowledge.

The Measuring Linguistic Complexity (MLC) shared task aims to bring together different measures of linguistic complexity, encouraging the use of Universal Dependencies (UD) treebanks (Nivre et al., 2016). The UD project defines a unified tagset, and the UD treebanks already include a large number of languages.[1] The multi-lingual focus of the UD project requires paying attention to linguistic typology (Croft, Nordquist, Looney, & Regan, 2017), and the treebanks, in return, constitute a promising resource for the typological (and in general multi-lingual) research. Not surprisingly, the MLC shared task offers a subset of the UD treebanks as the data set for measuring complexity of (subsystems of) languages.

In this paper, we present a number of quantitative measures of morphosyntactic complexity, namely, *type/token ratio* (TTR, e.g., Kettunen, 2014); *mean size of paradigm* (MSP Xanthos et al., 2011); *entropy of morphological-feature distribution*; *entropy of morphological-feature distribution conditioned on the word*

[1]Current UD release (v2.1) includes over 100 treebanks covering 64 languages. The candidate treebanks for the upcoming release includes treebanks for 16 other languages.

forms; *entropy of word-form distribution conditioned on morphological features*; *and part-of-speech tag n-gram perplexity*, calculated on the MLC selection of the 37 UD treebanks.

2. Measures

We report five measures (TTR, MSP, and variants of morphological feature entropy) for measuring morphological complexity, and one, POS tag n-gram perplexity, for measuring syntactic complexity. Except the first two (TTR and MSP), the measures discussed here are all suitable for richly-annotated corpora, and to our knowledge not used in this form in the previous literature.

2.1. *Type/token ratio (TTR)*

The TTR is a time-tested metric for measuring linguistic complexity. When used as a measure of complexity of a language, high TTR indicates rich morphology. Since the TTR depends on corpus length, it is a common practice to calculate the TTR using a fixed window size (Kettunen, 2014). We calculate the TTR on a fixed-length random sample, and take average over multiple samples. The sampling procedure is described in Section 3.

2.2. *Mean size of paradigm (MSP)*

Xanthos et al. (2011) propose the MSP as a measure of morphological complexity, and show its relation with the acquisition of morphology by young learners. The MSP is simply the number of word-form types divided by the number of lemma types. The MSP also depends on the text size. Hence, similar to Xanthos et al. (2011), we use a sampling-based approach (as in the TTR calculation).

2.3. *Morphological feature entropy (MFE)*

Any corpus that annotates words (or tokens) with a set of labels defines a categorical distribution. With MFE (defined in Equation 1), we estimate the categorical distribution of morphological features from the treebank, and calculate its entropy.

$$\mathrm{MFE} = -\sum_{\mathrm{f}} \mathrm{p}(\mathrm{f}) \log_2 \mathrm{p}(\mathrm{f}) \tag{1}$$

where f ranges over all observed feature-value pairs (e.g., `Case=Acc`) in the treebank. The probabilities are estimated with the maximum likelihood estimation (MLE) over all tokens (not types).

Intuitively, the entropy of this distribution indicates the richness of the morphological features encoded in the language. Everything being equal, a language with a larger morphological feature inventory will have higher MFE. However, the shape of the distribution also matters. A distribution that tends towards the uniform distribution, where all labels are equally likely, will also have higher entropy

compared to distributions that favor only a few high-probability (or frequent) features. Since the MFE does not depend on corpus size, we report values that are calculated over the complete available corpus.[2] This measure is similar to the *enumerative complexity* as defined by Ackerman and Malouf (2013).

2.4. *Conditional feature entropy*

Another aspect or dimension of morphological complexity is about transparency of a morphological system. Arguably, if we can predict morphological features from surface forms, and surface forms from morphological features, the language exhibits less complexity – e.g., when viewed from a learner's perspective.

As a first approximation for measuring transparency of the morphological system, we calculate two average conditional feature entropy values. The conditional entropy of a distribution Y given another distribution X is defined as

$$H(Y|X) = \sum_{x \in X, y \in Y} p(x, y) \log_2 p(y|x) \ .$$

The first measure we present, $\mathrm{CFE}_{w|m}$, is simply the conditional entropy of word forms given morphological features, $H(w \mid m)$, and the second measure, $\mathrm{CFE}_{m|w}$, is the conditional entropy of features given word forms, $H(m \mid w)$. It should be noted that these measures do not only measure the complexity of the morphological system but also measure the lexical complexity or ambiguity.

The conditional entropy measures we use are similar to *integrative complexity* defined by Ackerman and Malouf (2013). However, our measures reflect actual usage as reflected by the morphologically annotated corpora at hand, as opposed to the paradigm tables extracted from descriptive grammars.

2.5. *POS tag n-gram perplexity (POSP)*

As a measure of predictability of strictness of word order, we also compute the average perplexity of the UD POS tag n-grams. The perplexity is a popular measure of unpredictability in computational linguistics literature. It is defined as $2^{H(X)}$, where $H(X)$ is the entropy of a probability distribution X (of POS tag sequences in our case). The intuitive interpretation of POSP is the average number of possible POS tags after each position in the corpus. Intuitively, the languages with more strict word order is expected to have lower entropy (hence lower POSP). The POSP should correlate with the morphological complexity, particularly MFE, since rich morphology is typically associated with flexibility in the word order.

In this paper, we only present results of bigram perplexity. However, this can easily be extended to use higher order n-grams, or using entropy rate (Kontoyiannis, Algoet, Suhov, & Wyner, 1998; Gao, Kontoyiannis, & Bienenstock, 2008) for estimating the entropy of the POS tag sequence.

[2] However, the estimation of the underlying distribution will be better with larger corpora.

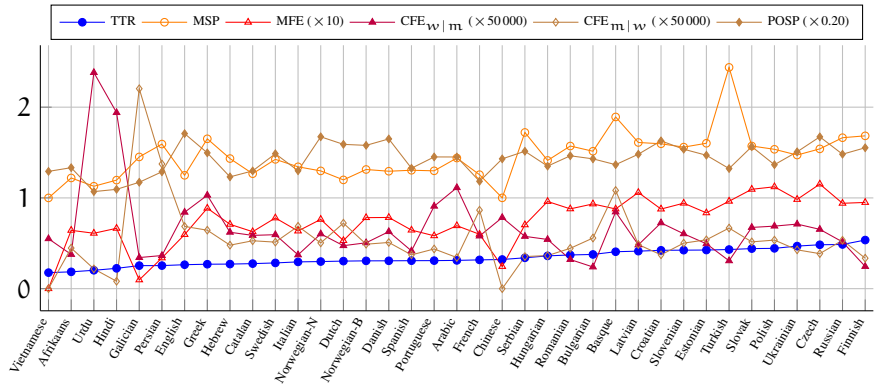

Figure 1. The values of the complexity measures. The measures are linearly scaled to fit into the same y-axis range, the languages are sorted in order of increasing TTR.

3. Data and experimental setup

The data set contains 37 treebanks from Universal Dependencies (UD) project, from 36 languages.[3] Although all treebanks conform to UD v2 annotation scheme, the sizes of the treebanks and some aspects of annotations vary considerably. The smallest treebank (Hungarian) has 1 801 sentences and 42 032 tokens, and the largest (Czech) consists of 87 914 sentences and 1 506 484 tokens. All treebanks, except Galician, include morphological feature annotations. The usage of UD POS tag inventory is relatively stable across languages. The number of POS tags used vary between 14 and 18. The morphological features and relation types used in different treebanks are more varied, ranging between 2 to 29 and 25 to 55 for morphological feature labels and dependency labels, respectively.

As noted above, some of our measures depend on text size. To be able to get comparable measures, we calculate TTR and MSP from 20 000 tokens sampled randomly. The numbers we report are the mean of 1 000 random samples.[4]

4. Results and Discussion

We present values of all measures discussed in Figure 1. The correlation between the languages are reported in Table 1. The overall results agree with our expectations and the earlier literature. The languages known to be more morphologically complex, are placed on the upper end of the scale with respect to measures that indicate enumerative morphological complexity. However, we also observe that

[3]Norwegian is represented by two treebanks, with different, but closely related dialects that also follow different orthographic conventions.

[4]The source code used for calculating the measures is publicly available at `https://github.com/coltekin/mlc2018`.

Table 1. Correlations between all measures. The values presented in the upper triangle matrix are Pearson's correlation coefficient, while Spearman's rank correlation is listed in the lower triangle.

	TTR	MSP	MFE	$\text{CFE}_{w\|\text{m}}$	$\text{CFE}_{\text{m}\|w}$	POSP
TTR		0.617 3	0.740 2	–0.333 5	–0.064 5	0.460 1
MSP	0.651 5		0.556 9	–0.308 9	0.276 4	0.142 0
MFE	0.776 4	0.658 4		–0.063 1	–0.237 1	0.418 5
$\text{CFE}_{w\|\text{m}}$	–0.100 8	–0.235 4	–0.027 3		–0.315 6	–0.337 7
$\text{CFE}_{\text{m}\|w}$	–0.027 3	0.254 5	–0.029 9	–0.292 3		–0.175 3
POSP	0.422 2	0.236 8	0.402 3	0.122 3	–0.022 3	

there is a modest but negative correlation between the enumerative complexity and integrative complexity measures used in this study. Furthermore, the (enumerative) morphological complexity, as expected, is also moderately correlated with flexibility of the word-order of the language measured by POSP.

The results also show some curious differences, e.g., Chinese showing moderately high TTR, despite lower MSP and MFE. Some of these, e.g., unexpectedly low MFE for Galician, however, is due to lack of annotations in the particular treebank. POSP seems to correlate with morphological complexity measures, indicating that POS tag sequences are less predictable in morphologically rich languages. However, some observations in Figure 1 needs further investigations. For example, the fact that Germanic languages, including English, showing rather high POSP, and despite being morphologically complex, Turkish showing showing a low POSP. Some of these differences may be due to the fact that our measurements are based on bigrams, hence being sensitive word order flexibility in local contexts, e.g., noun phrases, rather than flexibility at the level of the clause.

There are two major differences between the current study (also many others in this volume) and most earlier corpus- and grammar-based work on quantifying linguistic complexity. First, we make use of rich linguistic annotations, which offer many novel ways to measure linguistic complexity. Second, unlike many earlier studies, our material is not a (translated) parallel corpus collection. This allows measuring the complexity on a more 'natural' linguistic data, however, it also requires measures that indicate the differences between the languages, rather than other dimensions such as domain, genre or style. Compared to works that are based on descriptive grammars, working with relatively small corpora may result in missing some (rare) linguistic constructions. In this respect, larger (automatically annotated) data sets can be useful, or recent grammar-book treebanks (Çöltekin, 2015; Rama & Vajjala, 2017) may offer an interesting middle ground.

Although the measures and the results presented here needs further investigation and refinements that are beyond the scope of this short paper, the results are encouraging about using richly and uniformly annotated corpora, such as UD treebanks, for investigating many aspects of linguistic complexity.

References

Ackerman, F., & Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture. *Language*, *89*(3), 429–464.

Bentz, C., Alikaniotis, D., Cysouw, M., & Ferrer-i Cancho, R. (2017). The entropy of words—learnability and expressivity across more than 1000 languages. *Entropy*, *19*(6), 275.

Çöltekin, Ç. (2015). A grammar-book treebank of Turkish. In M. Dickinson, E. Hinrichs, A. Patejuk, & A. Przepiórkowski (Eds.), *Proceedings of the 14th workshop on treebanks and linguistic theories (TLT 14)* (pp. 35–49). Warsaw, Poland.

Croft, W., Nordquist, D. I., Looney, K., & Regan, M. (2017). Linguistic typology meets universal dependencies. In *Proceedings of the 15th workshop on treebanks and linguistic theories (TLT15)* (pp. 63–75). Bloomington, IN, USA.

Dahl, Ö. (2004). *The growth and maintenance of linguistic complexity*. John Benjamins.

Gao, Y., Kontoyiannis, I., & Bienenstock, E. (2008). Estimating the entropy of binary time series: Methodology, some theory and a simulation study. *Entropy*, *10*(2), 71–99.

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, *5*(3), 206–213. doi: 10.1080/09296179808590128

Kettunen, K. (2014). Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, *21*(3), 223-245. doi: 10.1080/09296174.2014.911506

Kontoyiannis, I., Algoet, P. H., Suhov, Y. M., & Wyner, A. J. (1998). Nonparametric entropy estimation for stationary processes and random fields, with applications to english text. *IEEE Transactions on Information Theory*, *44*(3), 1319–1327.

Koplenig, A., Meyer, P., Wolfer, S., & Mueller-Spitzer, C. (2017). The statistical trade-off between word order and word structure–large-scale evidence for the principle of least effort. *PloS one*, *12*(3), e0173614.

Newmeyer, F. J., & Preston, L. B. (2014). *Measuring grammatical complexity*. Oxford University Press.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., ... Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 23–28). Portorož, Slovenia.

Rama, T., & Vajjala, S. (2017). A Telugu treebank based on a grammar book. In *Proceedings of the 16th international workshop on treebanks and linguistic theories* (pp. 119–128). Prague, Czechia.

Stump, G. (2017). The nature and dimensions of complexity in morphology. *Annual Review of Linguistics*, *3*, 65–83.

Xanthos, A., Laaha, S., Gillis, S., Stephany, U., Aksu-Koç, A., Christofidou, A., … Dressler, W. U. (2011). On the role of morphological richness in the early development of noun and verb inflection. *First Language*, *31*(4), 461-479. doi: 10.1177/0142723711409976

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.

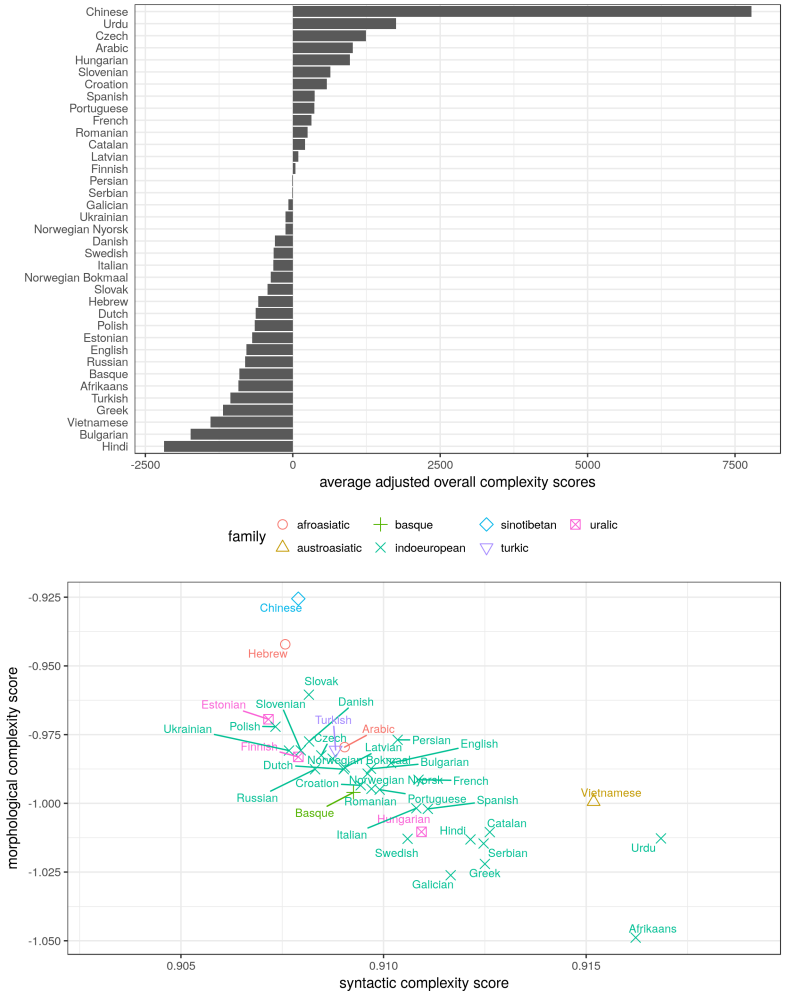

Figure 1. Upper plot: Overall complexity hierarchy. Negative residuals indicate below-average complexity; positive residuals indicate above-average complexity. Lower plot: Morphological by syntactic complexity. Abscissa indexes increased syntactic complexity; ordinate indexes increased morphological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

CONTRASTING PHONETIC COMPLEXITY ACROSS LANGUAGES: TWO APPROACHES

Caleb Everett [*1]

[*]caleb@miami.edu
[1]Anthropology, Psychology, University of Miami, USA

This paper examines phonetic complexity via two approaches that rely on transcribed word lists. Both approaches focus on results obtained for a 37-language sample, but contrast these results with findings from about 7000 language varieties. For the first approach, complexity is measured simply as the ratio of types of phones to tokens of phones, for each list representing a particular language variety. The second approach operationalizes complexity as unpredictability of sound usage, and simplicity as predictability. Predictability is based on the global mean frequency of occurrence of 41 sound types across all language varieties in the data. These global frequencies are then used to predict sound usage in the 37 languages focused upon, with less predictable languages deemed more "complex". Three languages in the sample are found to be complex according to both metrics explored here, while two languages are found to be simple according to both metrics. These findings are exploratory given the limitations of the word lists tested.

1. Introduction

Languages vary markedly in terms of the number of sounds they utilize. One could argue that languages with more phonemes represent complex phonological systems, though such a claim overlooks non-phonemic parameters including syllable structure and prosodic phenomena. Still, we can speak of specific kinds of complexity, e.g. complexity of phonemic inventories, without making presumptions regarding overall phonological, phonetic, or otherwise linguistic complexity. In this study I offer two approaches to looking at the complexity of languages' variant usage of sounds, both of which focus upon the phonetic units in basic transcriptions of 40-100 words (Swadesh-type lists). I apply both methods to the 37-language sample but, as critical background to this sample, I also apply the metrics to thousands of other languages.

2. Type:token ratio of phonetic segments

The first metric of complexity is simply the type:token ratio of transcribed phonetic units. The assumption underlying this metric is that languages with a greater density of sound types are more complex in terms of their sound-type inventories. I say "sound-types" as opposed to phonemes because this study relies on the ASJP database, a collection of roughly 7000 word lists that are phonetically transcribed. The phonetic transcriptions in the database are somewhat broad, as they use 41 basic sound types (Wichmann et al. 2016). Still, despite any limitations, there are advantages to using a database representing so many languages, as we can contrast our results for the 37-language sample with results from the bulk of the world's languages. (Over 4500 distinct ISO codes are represented in the data.)

To calculate the type:token ratio, I simply summed the number of unique sound types represented in a word list, and then divided that sum by the total number of sound tokens represented in the list. Secondary symbols for nasalization and other phenomena were ignored. Since this study focused on phonetic segments as opposed to phonemes, two-sound sequences such as prenasalized stops were treated as separate sounds. To contextualize the type:token ratios obtained for the 37-language sample, I gathered type:token ratios for about 7000 other varieties in the database. (I excluded varieties for artificially constructed languages.) I then obtained family-level averages of these ratios. The 264 family groupings were based on the WALS database (Dryer et al. 2013). Family means of type:token ratios ranged from 0.026 to 0.283. The overall mean across families was 0.121. The mean for the 37-language sample was about the same, at 0.119. (For a list of all family means, see the supplemental material.) The following ordering was observed, for the 37-language sample, from highest to lowest type:token ratio: 1. Norwegian (Nynorsk) 2. Catalan 3. Portuguese 4. Afrikaans 5. Danish 6. Arabic 7. Swedish 8. Polish 9. Czech 10. Slovak 11. Slovenian 12. Urdu 13. Turkish 14. Hebrew 15. Dutch 16. Galician 17. Croatian 18. Romanian 19. Italian 20. Norwegian (Bokmaal) 21. Bulgarian 22. Ukrainian 23. Vietnamese 24. Latvian 25. Mandarin 26. Greek 27. English 28. Hungarian 29. French 30. Persian 31. Hindi 32. Estonian 33. Russian 34. Serbian 35. Finnish 36. Spanish 37. Basque (See results file.)

To be clear, the suggestion being made here is not that languages with higher type:token ratios are necessarily more complex in terms of articulation. I am simply proffering one way of exploring phonetic segment complexity, one that

could be tested for associations with socioecological factors. This approach could also be applied to more robust intra-linguistic samples.

3. Predictability of sounds' usage rates

Another way to think of phonetic complexity is in terms of deviation from a typologically based expectation of languages' usage of individual sound types. According to such an approach, languages that use crosslinguistically uncommon sounds frequently, or common sounds very infrequently, would be more unpredictable and therefore more "complex" in a typological sense.

Given the lists of sounds in a particular word list, we can predict (roughly) how much each sound is used (Everett, under revision). For instance, we may predict that an alveolar nasal is used frequently, a voiceless alveolar stop a bit less so, a voiced alveolar stop even less, and so on. (Assuming these sounds are all present in the language in question.) The second metric for complexity adopted here relies on the fact that sounds' "usage rates" are somewhat predictable. Usage rates refer to the proportion of all the sound tokens in a given word list that are represented by a given sound. For instance, if there are four tokens of [t] in word list, out of 400 total sounds in the words in the list, then the usage rate of [t] is simply 0.01. Usage rates can be used to test the predictability of the occurrence of sounds across the world's languages. To do so, I adopted the following five steps: 1) Usage rates were obtained for all 41 sounds in each of the 6902 language varieties tested. 2) The average family-level usage rates were found for all sounds for each of 264 WALS language families. 3) These family-level averages were then averaged, resulting in phylogenetically controlled average usage rates for all sounds. 4) The sounds were then ranked according to their usage rates, at a global scale. (Sound rankings and mean usage rates are presented in the supplemental material.) 5) These global rankings were used to generate the predicted usage rates of sound types for individual languages, and these predicted usage rates were then contrasted with actual usage rates. Step 5 requires some elaboration: How are sound rankings, from most (#1) to least (#41) used in the world's languages, transformed into predicted usage rates? I transformed the rankings into predicted usage rates via the Borodovsky and Gusein-Zade formula. This formula was developed to predict the frequency of phonemes within a language from the frequency ranking of phonemes for that language (Tambovtsev and Martindale 2007). The formula allows us to predict a phoneme's intralinguistic frequency (f_r) from its intralinguistic rank *(r)*:

$$f_r = \frac{1}{n}\,(log(n+1) - log\ r)$$

For this study, I used the same formula but used the crosslinguistic rank, arrived at in step 4 above, as *r*. For *n*, I used the number of basic sounds in the database, or 41. (I should note that this formula only provides a marginal improvement to simply using the observed global average usage rates as the expected usage rates within each language, without using any transformation at all.)

With the predicted usage rates of the 41 sounds in hand, I then focused on the actual usage rates of the sounds in the 37-language sample. The association between predicted and actual usage rates, for all 37 varieties, is depicted in Figure 1. For each of the 37 languages, I ran a regression testing the association between predicted usage and actual usage. Higher R^2 values correspond to greater overall predictability. R^2 values ranged from 0.76 to 0.15, with a median of 0.51. Lower R^2 values are suggestive of greater usage-based deviance from a crosslinguistc norm, a kind of typologically-based complexity. The association is quite robust in most cases, but some varieties are clearly more predictable vis-à-vis their usage of sounds in these word lists. (See Figure 1 below. See results file for R^2 values of each of the 37 languages.)

4. Conclusion

I have outlined two potential approaches, of many, to measuring phonetic complexity. Each approach is based on a different interpretation of what is meant by complexity. One considers languages with predictable usage rates to be less complex (though admittedly this operationalization equates typologically anomalous usage with complexity, a strategy open to debate), the other considers languages that rely repeatedly on the same sounds, with relatively sparse usage of distinct sound types, to be less complex. The metrics resulting from these approaches are admittedly coarse but, I think, useful as exploratory measures. The complexity rankings of the 37 languages are somewhat similar for both metrics (Spearman's rho=0.44, p=.007). Finally, some remarks on individual languages: Interestingly, three closely related languages are the three most complex languages according to the predictability metric: Swedish, Danish, and Norwegian (Bokmaal), in that order. Swedish and Danish are also in the top 7 according to the type:token metric. Norwegian (Bokmaal) is not amongst the most complex according to the type:token metric, though Norwegian (Nynorsk) is. So there is some Scandinavian flavor to the more complex varieties according to both metrics, but also some cross-dialectal variability (which admittedly may simply be the artifact of the small sample

sizes). In contrast, both Finnish and Basque are ranked amongst the three "least complex", for both metrics. Of course it remains to be seen just how much these findings, based as they are on short word lists, are representative of larger patterns in these languages. What we can say is that, given the metrics and data utilized here, there is some observable though modest coherence at both ends of the range of complexity, for this sample of 37 languages.

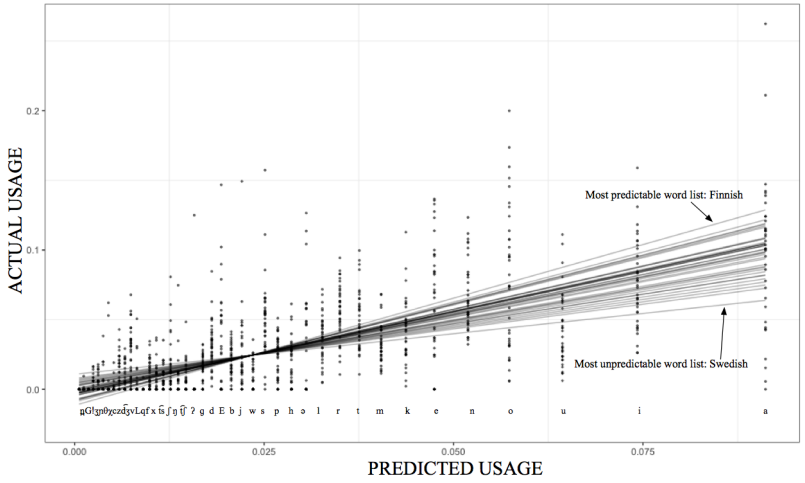

Figure 1. Relationship between predicted usage and actual usage, for each of the 37 languages in the sample. Each LMS line depicts the association between the typologically based predicted usage of 41 sound types and a language's actual usage of those sound types (judging from the transcribed word lists). Each column of 37 dots represents the usage rates of one of the 41 sounds, across each language in the 37-language sample.

References

Everett, C. (Under revision) The predictability of sound usage across languages.

Dryer, Matthew S. & Haspelmath, Martin (eds.) (2013) *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Tambovtsev, Y. & Martindale, C. (2007) Phoneme frequencies follow a Yule distribution: The form of the phonemic distribution in world languages. *SKASE Journal of Theoretical Linguistics* 4.2

Wichmann, S., Holman, E., and Brown, C. (2016) The ASJP Database.

POS TAG PERPLEXITY AS A MEASURE OF SYNTACTIC COMPLEXITY

Kilu von Prince[*1,2] and Vera Demberg[2]

[*]Corresponding Author: kilu.von.prince@hu-berlin.de
[1]Humboldt-Universität, Berlin, Germany
[2]Universität des Saarlandes, Saarbrücken, Germany

Comparing languages of the world with respect to their complexity is a long-standing open question in linguistics. We here focus on syntactic complexity, a concept that has been particularly hard to address due to the lack of readily available syntactically annotated corpora and the intricacies of syntactic theories. We propose to use a simple information-theoretic measure, perplexity, on the POS tag sequence of texts. Perplexity captures how predictable POS tags are on average given their recent co-texts. Calculating perplexity based on POS tag sequences helps us to abstract away from morphological or lexical features of the language, in order to get at the predictability of word order. In this paper, we compare POS tag perplexity to other recently proposed measures of syntactic complexity, and evaluate measures by correlating them with expert-proposed scores of syntactic flexibility (Bakker 1998).

1. Introduction

The question of how and why languages may differ in terms of their overall or partial complexity is one of the oldest and most hotly debated issues in typology (Nichols, 1992; Trudgill, 2011; McWhorter, 2001; Sampson, 2009; Joseph & Newmeyer, 2012). Since Juola (1998), several attempts have been made to assess the complexity of languages based on texts rather than typological features (Juola, 2008; Futrell, Mahowald, & Gibson, 2015; Ehret & Szmrecsanyi, 2016; Bentz, 2016; Koplenig, Meyer, Wolfer, & Müller-Spitzer, 2017). In this paper, we will assess the viability of using trigram perplexity at the POS level for assessing cross-linguistic variation between non-parallel corpora. This measure is defined as below:

(1) Trigram perplexity:
$2^{-\frac{1}{N}\sum_{n=1}^{N} P(pos_n|pos_{n-2},pos_{n-1}) \log_2 P(pos_n|pos_{n-2},pos_{n-1})}$

The reason for working on POS tag sequences instead of words directly is to avoid possible confounds due to writing systems, lexical richness or choice, and, to some extent, compensate for the fact that the data we are

working with are not parallel texts.

2. Related Work

We compare the estimates of our method to various previously proposed measures, which we will briefly introduce here.
Expert-ratings of syntactic complexity. Bakker (1998) rated syntactic flexibility, consistency and consequence of languages based on twelve binary grammatical features as described in descriptive accounts and expert questionnaires on each language. Bakker's syntactic flexibility measure, which was also used in prior evaluations such as Ehret and Szmrecsanyi (2016) seemed like the most representative source of expert syntactic complexity ratings against which automatic measures can be evaluated.
Zip compression as an approximation to Kolmogorov complexity. Zip compression has been known to approximate Kolmogorov complexity and has previously been used as a measure of linguistic complexity (Juola, 1998, 2008; Ehret & Szmrecsanyi, 2016). We calculated zip compression for each of the corpora, to achieve best possible comparability with our proposed POS tag perplexity metric.
Ehret and Szmrecsanyi (2016). The authors used a parallel corpus consisting of translations of Alice in Wonderland into nine lanuages, compiled by Annemarie Verkerk, and non-parallel newspaper corpora from the same languages. To measure syntactic complexity, they masked syntactic regularities by randomly deleting 10% of all word tokens. They then measured the difference between the zip-compressed original text and the zip-compressed masked version.
Koplenig et al. (2017) used the massive Parallel Bible Corpus Mayer and Cysouw (2014), with translation into almost 1200 languages. They also created syntactically masked versions of each text by scrambling sentence-internal word order. They then measured an approximation to entropy for the masked and unmasked versions and calculated the difference between them as a measure of syntactic complexity.

3. Methods

We first analyzed the distribution of POS tags across corpora to ensure comparability. While some of those differences may reflect genuine cross-linguistic variation that speaks to differences in the size of syntactic inventories, we need to keep in mind that some variation may have been caused by language-external factors instead. For example, the Chinese corpus does not use the tag for subordinating conjunctions, even though it has very straightforward candidates for this category, such as yīnwéi, "because", which is tagged as an adposition instead. The Arabic corpus contains a much larger set of Other tags compared to the other corpora,

which may be an indication of the limits of the applicability of a universal tag set. Such inconsistencies are a potential cause for concern for future fine-grained comparative work, especially if the list of languages in the set expands to include more non-European languages.

To assess possible effects of POS tag distributions across languages, we also measured unigram perplexity $2^{-\sum_{pos \in \mathrm{POS}} P(pos) \log_2 P(pos)}$; the probability of a POS tag pos was estimated in terms of its frequency in the corpus. Unigram perplexity over POS tags hence quantifies the differences in entropy of the POS tag inventory of a language. In order to separate out trigram perplexity from unigram perplexity, we propose an additional measure: trigram perplexity divided by unigram perplexity to quantify the predictability of syntactic categories given previous context compared to POS tag frequencies. This measure gives us a sense of the predictability of word order that is independent from how big and balanced the inventory of POS tags is.

For calculating perplexity and zip compression, we extracted POS tags from each of the corpora, split the files into chunks of 42k tags – the size of the smallest corpus. This allowed us to also assess the effect of corpus size on complexity scores and also allowed us to calculate variance for different subsets of texts for the same language. We found that estimates were generally reliable; our results below report perplexities for the complete dataset for each language, as our experiments showed that estimates on 42k subcorpora correlated at Spearman's rho 0.98 with estimates from the full corpora. This result demonstrates that working on POS sequences avoids having to deal with data sparsity issues.

4. Results

Figure 1 shows that our perplexity measures are correlated with the syntactic flexibility values proposed by Bakker (1998) ($\rho = .45$, $p < 0.05$). The statistical analysis also shows that our measures predict human ratings by Bakker more reliably than previously proposed measures (Juola, 1998; Ehret & Szmrecsanyi, 2016; Koplenig et al., 2017). Figure 1 visually illustrates the correlation between our Trigram/Unigram measure and Bakker flexibility scores.

Among the languages for which there are no Bakker scores, our perplexity measures would predict that Hebrew, Afrikaans, Hindi, Urdu and Arabic are among the syntactically more complex languages (if complexity means lack of flexibility). For Vietnamese, Chinese, Persian, Hungarian, Ukrainian and Czech, the classification as mid range or low syntactically complex languages depends on whether unigram perplexity is taken into account: Ukrainian and Czech have high unigram perplexity, and would hence be classified as highly flexible languages in the trigram measure, but

Table 1. Spearman's correlation between various automatic measures of syntactic complexity and Bakker (1998) flexibility scores. Column dir indicates whether the correlation with Bakker scores is expected to be positive or negative. The right-hand part of the table compares only the set of seven languages used both as part of Ehret & Szmrecsanyi (2016) and as part of the datasets provided for the present workshop.

measure	dir	corr	pval	# lang	corr	pval	# lang
avgzip	neg	-0.32	0.11	26	-0.71	0.07	7
Koplenig et al. '17	neg	-0.36	0.08	24	-0.29	0.53	7
E&S'16: Parallel Alice	neg				-0.71	0.07	7
E&S'16: News	neg				-0.43	0.33	7
unigram perplexity	NA	0.12	0.57	26	-0.09	0.85	7
trigram perplexity	pos	0.45	0.02	26	0.87	0.01	7
trigram/unigram	pos	0.44	0.02	26	0.85	0.01	7

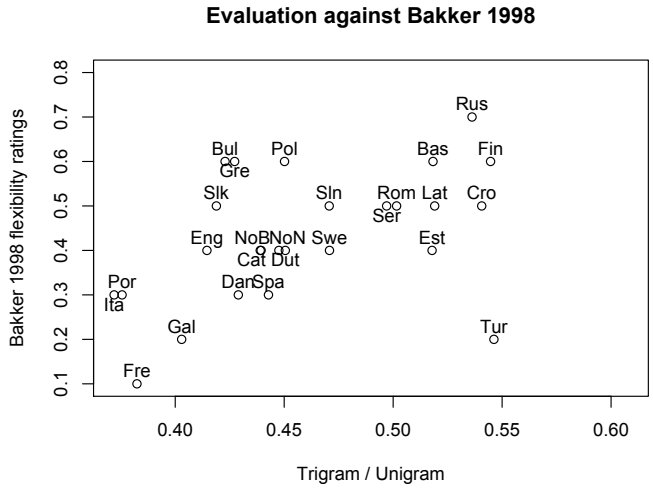

Figure 1. Correlations of trigram perplexities divided by unigram perplexities with flexibility values in Bakker 1998.

as medium complexity languages in the trigram/unigram measure. On the other hand, Persian, Vietnamese and Chinese have low unigram perplexities and hence are only classified as highly flexible languages in the trigram/unigram measure.

5. Discussion

In sum, our results show that surprisal values of POS tags, even in relatively small, non-parallel corpora can be a meaningful measure of syntactic complexity and perform better than similar methods at the word level. Our

hypothesis is that measures at the word level may be compromised by unrelated factors, such as the rate of homographs in a given corpus. By focusing on POS levels, these factors can be avoided.

While we did find a significant overall correlation of our methods with Bakker (1998), some languages show a much better fit than others. In particular, Turkish, Bulgarian and Greek are outside the expected range. At this point, we do not have a perfect explanation for these mismatches. Unfortunately, the ratings in Bakker (1998) are not entirely transparent: it is unclear exactly which feature combination is assigned to each language. Despite the very low score of 0.2 (on a scale from zero to one) from Bakker (1998), Turkish could be expected to receive a relatively high score, since it is well-known to be a free-word-order language. In this case, it may therefore be that the corpus-based measure gives a more accurate estimate of the actual flexibility the language has. Bulgarian and Greek are also known for their relative freedom of word order, in line with Bakker's high scores, so their comparatively low values of POS trigram perplexity are rather unexpected. It might be that these differences between description-based assessments and corpus-based measures speak to actual differences between theoretical possibilities and their implementation in language use. It is however also possible that these mismatches are due to imperfections in the POS tagger (which may affect some languages more than others) or a non-representative selection of syntactic features in Bakker (1998).

References

Bakker, D. (1998). Flexibility and consistency in word order patterns in the languages of europe. In A. Siewierska (Ed.), Constituent order in the languages of Europe: Empirical approaches to language typology (Vol. 20, p. 383-420). Mouton De Gruyter.

Bentz, C. (2016). The low-complexity-belt: evidence for large-scale language contact in human prehistory? In S. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Feher, & V. T. (Eds.), The evolution of language: Proceedings of the 11th international conference.

Ehret, K., & Szmrecsanyi, B. (2016). An information-theoretic appraoch to assess linguistic complexity. In R. Baechler & G. Seiler (Eds.), Complexity, isolation and variation (p. 71-94). Berlin, New York: de Gruyter.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. Proceedings of the National Academy of Sciences, 112(33), 10336-10341.

Joseph, J. E., & Newmeyer, F. J. (2012). All languages are equally complex: The rise and fall of a consensus. Historiographia Linguistica, 39, 341-368.

Juola, P. (1998). Measuring linguistic complexity: The morphological tier.

Journal of Quantitative Linguistics, 5, 206-250.
Juola, P. (2008). Assessing linguistic complexity. In M. Miestamo, K. Sinnemäki, & F. Karlsson (Eds.), Language complexity – typology, contact, change (Vol. 94, p. 89-108). Amsterdam, Philadelphia: John Benjamins Publishing Company.
Koplenig, A., Meyer, P., Wolfer, S., & Müller-Spitzer, C. (2017). The statistical trade-off between word order and word structure – large-scale evidence for the principle of least effort. PloS one, 12(3).
Mayer, T., & Cysouw, M. (2014). Creating a massively parallel bible corpus. In K. Choukri, D. T., L. H., B. Maegaard, & J. Mariani (Eds.), Proceedings of the ninth international conference on language resources and evaluation (p. 26-31). Reykjavik, Iceland: European Language Resources Association (ELRA).
McWhorter, J. H. (2001). The world's simplest grammars are creole grammars. Linguistic Typology, 5, 125-166.
Nichols, J. (1992). Linguistic diversity in space and time. Chicago and London: University of Chicago Press.
Sampson, G. (2009). A linguistic axiom challenged. In G. Sampson, D. Gil, & P. Trudgill (Eds.), Language complexity as an evolving variable (p. 1-18). Oxford, New York: Oxford University Press.
Trudgill, P. (2011). Sociolinguistic typology: Social determinants of linguistic complexity. Oxford University Press.

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.

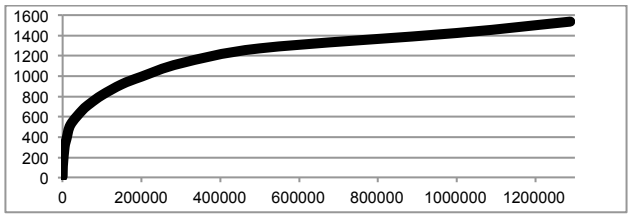

Figure 1. Cumulative unique dependency relationships per total number in corpus (Czech).

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.

MORPHOSEMANTIC COMPLEXITY

Bill Thompson[*1] and Gary Lupyan[2, 1]

[*]Corresponding Author: biltho@mpi.nl
[1]Language and Cognition Department, Max Planck Institute for Psycholinguistics
[2]Department of Psychology, University of Wisconsin-Madison

We describe *morphosemantic complexity*, a new measure of morphological complexity based on traversal of semantic space. Imagine meaning as a multi-dimensional space and the transition from lemma to wordform as a direction in this space. We propose a formulation of morphological complexity as the variability among these traversals. As an example, consider the English past-tense as the collection of difference vectors between lemmas and their inflected forms. A past-tense paradigm showing a high degree of semantic regularity is one in which the traversal from "walk" $\rightarrow$ "walked" has a similar direction as the traversal from "feel" $\rightarrow$ "felt" and "is" $\rightarrow$ "was". That is, the variance between these difference vectors is small. On our measure, the fact that some English words (e.g. "feel"/"felt", "is"/"was") violate the usual English past-tense pattern is not relevant. Rather, our measure picks up on the semantic "consistency" of inflectional paradigms. Our results show that measuring morphological complexity in this way provides strong correlations with corpus-based measures such as C_{WALS} (Bentz, Ruzsics, Koplenig, & Samardzic, 2016) and entropy-based $D_{structure}$ (Koplenig, Meyer, Wolfer, & Mueller-Spitzer, 2017), but appears to also account for unique variance, while offering additional advantages which we describe below.

1. Method and Rationale

We obtained word-embeddings for the 37 languages listed in this task. The embeddings are 300-dimensional vectors derived from training a Skipgram model on Wikipedia in each language. We used pretrained vectors made available by Facebook Artificial Intelligence Research (Bojanowski, Grave, Joulin, & Mikolov, 2016). These vectors have the property that similar vectors generally correspond to semantically similar words (Mikolov, Chen, Corrado, & Dean, 2013; Chen, Peterson, & Griffiths, 2017; Nematzadeh, Meylan, & Griffiths, 2017; Hollis & Westbury, 2016). Most relevant to our purposes is the ability to capture compositional aspects of word meaning via numerical operations on the word vectors. A canonical example is that the vector for "king" minus the vector for "man" plus the

vector “woman” puts us in part of the semantic space closest to “queen” (Mikolov et al., 2013). The vector operations can be applied to morphological transformations as well: the difference between “cats” and “cat”, added to “tree”, produces a vector most similar to the word “trees”. Importantly, this analogy-type process operates in semantic space rather than wordform space.

For each of the 37 languages, we obtained from the CoNLL-U annotations form-lemma pairs for every token in each datafile. For all form-lemma pairs for which we were able to obtain word vectors for both words, we subtracted the lemma vector from the base-word vector producing a difference vector. When form and lemma differ, the difference vector can be taken to represent the *meaning* of the morphological transformation. When the stem and lemma were identical, the difference vector is simply 0. Because our semantic vectors are linked to string representations of words, we cannot distinguish parts of speech; “rain” (N) and “rain” (V) would therefore be represented by the same vector.

Call the total collection of difference vectors for a given language its *difference-set*. In a morphologically simple language, the difference-set will be mostly vectors of zeros. As a result, we would expect less variance among vectors in the difference set, and less absolute semantic volume (i.e. average distance from zero). In a morphologically rich language, the difference-set will exhibit both more variance and volume. The distance and variance measures can also diverge. Figure 1 visualises these variables in three languages. Each arrow in these figures corresponds to a single difference vector, drawn very faintly. After projecting word vectors onto a two dimensional space, we plotted the angle and distance of the traversal from lemma to wordform. In English, relatively little semantic work in being done by morpholpogy (short arrows), and the traversals tend to cluster into a small number of similar categories (shown by arrows that appear dark, because they layer on top of eachother at similar angles). Turkish and Farsi (Persian) both do lots of semantic work with morphology (long arrows), but lower variance of angles in Farsi than Turkish suggests a a smaller number of semantic transformations.

We obtained the difference-set for all 37 languages and computed several measures:

- **Semantic Distance (All Tokens) & (Non-Identical Tokens)**: The total distance travelled between lemma and form (i.e. the sum of by-component squared distances from zero) vectors among all unique word pairs, including cases where lemma and form are the same word, or not, respectively. This measure quantifies the amount of semantic work being done by morphology.

- **Semantic Variance (All Tokens) & (Non-Identical Tokens)**: The variance among difference vectors for all unique word pairs, including cases where lemma and form are the same word, or not, respectively.

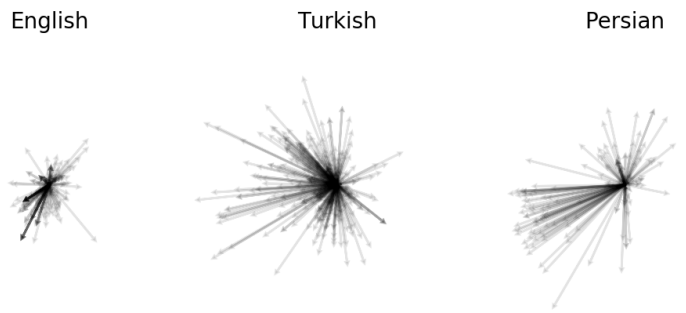

Figure 1. Distance and angle of all difference vectors (traversals between lemma and wordform) in three languages, projected into two-dimensional vector space and arranged around a common origin.

2. Results

The supplementary materials for this article contain a dataset which lists, for each language: the measures listed above plus C-WALS (Bentz et al., 2016) and D-structure (Koplenig et al., 2017). For completeness, we also include the following variables:

- **Lemma = Wordform Proportion (Tokens) & (Types)** — The proportion of all attested & all unique words respectively whose lemma matches the infected form.
- **Number of Morphological Categories** — The number of categories catalogued in the CoNLL-U files (e.g., Tense, Person, Aspect, Gender)
- **Morphological Sum** — The sum of the total values for each category, e.g., Feminine, Masculine, Past-tense, etc.
- **GZIP-R** — A measure of morphological complexity similar to $D_{structure}$ (Koplenig et al., 2017): [1-size of gzipped plain-text]/[size of gzipped with word-substituted text] where word-substituted text is created by replacing each word with a random number of characters drawn from the frequency distribution of characters in the language. This results in disrupting compression gains that are based on reusing codes for stems in morphologically derived words.

Figure 2 shows simple Pearson correlations between the variables. Several of these are worth highlighting: a) the number of categories is a rather bad predictor of all measures of morphological complexity because most of the languages in this sample share most morphological categories, differing only in the number of values per category; b) The proportion of word forms that are equal to their

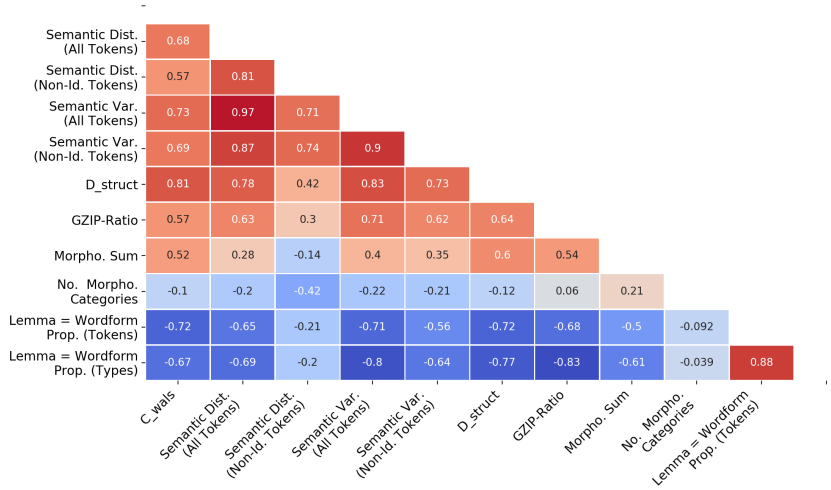

Figure 2. Correlation among our proposed measures, existing measures, and lower level morphological summary statistics.

lemmas (both as raw wordforms and proportion of unique wordforms) correlates to a surprising extent with previously published WALS-based measure (C_{WALS}) and entropy-based measures ($D_{structure}$), as well as our own entropy-based measure (GZIP-R); c) both our semantic distance and semantic variance measures are strongly correlated with C_{WALS}, $D_{structure}$ and GZIP-R. Table 1 shows a subset of these measures for the ten most and least complex langauges, as judged by our *Semantic Distance (All Tokens)* measure.

To check whether the high correlations between morphosemantics and existing complexity norms are confounded by variables such as *Lemma = Wordform*, we conducted a series of multiple regressions where these variables are partialed out. Details of these results are presented in the supplemantary materials. Both Semantic Distance (All Tokens) and Semantic Variance (All Tokens) are independently predictive of both C_{WALS} and $D_{structure}$, at significance levels $< .01$, even when controlling for the morphological measures we extracted from the CONLL-U parse.

As an initial test of the kind of small differences in complexity our semantic-distance measures is able to detect, we examined the closely-related languages Bokmål and Nynorsk (we also studied Serbian/Croatian, and found similar subtleties). Bokmål (lit. Book tongue) and Nynorsk (lit. New Norwegian) are two standardized forms of written Norwegian. Bokmål is more common, being used by about 87% of the population and, of the two varieties, has been strongly influ-

Language	Semantic Dist. (All Tokens)	Semantic Var. (All Tokens)	GZIP-R	D_struct	C_wals
Hebrew	35.54	24.86	0.24	0.52	0.53
Arabic	33.86	21.85	0.21	0.57	0.80
Persian	23.03	18.82	0.17	0.36	0.52
Turkish	21.14	20.44	0.22	0.60	0.78
Finnish	18.46	17.86	0.21	0.43	0.48
Estonian	17.89	17.20	0.17	0.41	0.62
Latvian	14.03	13.55	0.20	0.45	0.52
Serbian	13.49	12.98	0.17	0.37	0.44
Russian	12.96	12.30	0.27	0.42	0.45
Greek	12.93	12.20	0.22	0.32	0.45
⋮	⋮	⋮	⋮	⋮	⋮
Swedish	7.73	7.54	0.18	0.21	0.33
Italian	7.66	7.38	0.11	0.31	0.38
Portuguese	6.23	5.95	0.14	0.33	0.45
French	6.11	5.85	0.13	0.29	0.43
Danish	6.02	5.90	0.13	0.26	0.39
Catalan	5.83	5.62	0.13	0.35	0.23
Urdu	5.12	5.09	0.12	0.25	0.36
Dutch	4.87	4.81	0.13	0.27	0.33
Hindi	4.00	3.96	0.15	0.25	0.53
Afrikaans	3.89	3.84	0.13	0.19	0.12
English	3.47	3.41	0.10	0.19	0.33

enced by Danish. Nynorsk is a minority form used by 12.5% of Norwegians has resisted Danish influence to a greater extent. The treebanks for the two varieties are nearly the same size and show almost identical categories and values. Bokmål has two more values (reflexives and a passive voice) and so on this measure may be viewed as being slightly more complex (though the lack of reflexives and passive in Nynorsk appears to be an inconsistency in treebank coding). The greater complexity of Bokmål is also supported by Koplenig's entropy-based measure of structural complexity of bible translations ($D_{structure}$ Bokmål = .24; $D_{structure}$ Nynorsk=.22), as well as our own entropy-based estimate. In contrast, according to the morphosemantic complexity measure we compute here, Bokmål is simpler; it has lower semantic variance (i.e., having more semantically consistent morphological paradgims): Bokmål = 10.65, Nynorsk=14.24. Consistent with Bokmål being strongly influenced by Danish, its semantic variance is very close to that of Danish (10.81).

3. Future Directions

The work described here is preliminary. We are beginning to investigate whether it is possible to derive a similar measure from plain-text by sampling words in a corpus at a fixed edit-distances apart and computing their semantic distances, and variance among their distances. We are also investigating the use of morphosemantics to detect morphological paradigms without linguistic annotation, i.e., in a purely empirical way, by performing cluster-analysis of difference vectors.

References

Bentz, C., Ruzsics, T., Koplenig, A., & Samardzic, T. (2016). A comparison between morphological complexity measures: typological data vs. language corpora. In *Proceedings of the workshop on computational linguistics for linguistic complexity (cl4lc)* (pp. 142–153).

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Chen, D., Peterson, J. C., & Griffiths, T. L. (2017). Evaluating vector-space models of analogy. *arXiv preprint arXiv:1705.04416*.

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic bulletin & review*, *23*(6), 1744–1756.

Koplenig, A., Meyer, P., Wolfer, S., & Mueller-Spitzer, C. (2017). The statistical trade-off between word order and word structure–large-scale evidence for the principle of least effort. *PloS one*, *12*(3), e0173614.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nematzadeh, A., Meylan, S. C., & Griffiths, T. L. (2017). Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words. In *Proceedings of the 39th annual meeting of the cognitive science society.*

SYNTACTIC COMPLEXITY COMBINING DEPENDENCY LENGTH AND DEPENDENCY FLUX WEIGHT

CHUNXIAO YAN[*1], SYLVAIN KAHANE[1]

[*]Corresponding Author: yanchunxiao@yahoo.fr
[1]MoDyCo, Université Paris Nanterre & CNRS, Paris, France

1. Introduction

According to psychological research by Miller (1956), human memory is constrained to 7±2 elements. This cognitive constraint has an influence on natural languages at the level of syntax. Yngve (1960) proposed the notion of depth in his model, which makes it possible to model the grammar of the language by considering short-term memory. Nowadays, treebank resources make it possible to measure language performance on real data. Liu (2008) measured the dependency distance/length, which is the linear distance between a governor and a dependent, using dependency treebanks, and showed that the dependency distance has a tendency to be minimized (see also Tesnière, 1959: chapter 7; Futrell et al., 2015). Another typical question concerning the complexity of syntactic and cognitive ability concerns the limitation on the level of center-embedded constructions. Miller and Chomsky (1963) defined center-embedded constructions as a "nesting of dependencies, which occurs when X is embedded in another constituent Y, with material in Y to both the left and right of X," and remarked that increasing the levels of center-embedding makes the sentence incomprehensible. According to the psycholinguistic research by Lewis (1996), an English sentence has two levels of center-embedded clauses at most. In Japanese the total number can reach three (Lewis, 1996). This syntactic limitation is hypothesized to be related to the constraints of short-term memory. Kahane et al. (2017) considered the dependency flux, which is the set of dependencies linking a word on the left with a word on the right in a given position in the text and computed the flux weight, i.e. the maximum number of disjoint dependencies in the flux. As they showed, the flux weight, which measures the level of center-embedding constructions, is limited to 5 in the 70 treebanks of UD 2.0.

The dependency length carries only linear information and does not make it possible to measure the complexity of the configuration of dependencies, while the flux weight only evaluates the shape of the configuration of dependencies, without considering whether the dependencies in the configuration are long or short. We therefore propose a combined weight measure, in order to account for these two measurements at the same time. The calculation of dependency length, flux weight and combined weight will be presented in the next section.

2. Dependency flux

2.1. Flux size and flux weight

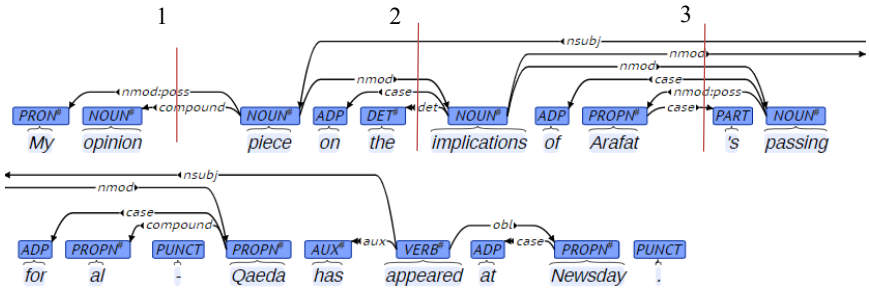

Figure 1. A dependency tree from UD-English-Original, with three positions considered

According to Kahane et al. (2017), dependency flux is the set of dependencies linking a word on the left with a word on the right in a given position. In Figure 1, three examples of flux positions are indicated by a vertical line: position 1 (opinion, piece), position 2 (the, implications), and position 3 (Arafat, 's). The flux size is the number of dependency links crossing the position. For position 1, we have two links, labeled *nmod:poss* and *compound*, that link a word to the left and a word to the right and the flux size is 2; for position 2, the flux size is 4; for position 3, the flux size is 6.

A set of dependencies is said to be disjoint if the dependencies do not share any vertex. The number of disjoint dependencies measures the center-embedding level (Kahane et al., 2017). For instance in position 2, there are two disjoint dependencies, [appeared -*nsubj*-> piece] and [implications -*case*-> on], which do not share any vertex, and represent exactly a center-embedded construction from the point of view of constituency analysis: [piece [on the implications] appeared]. In position 3, we find a set of four disjoint dependencies: [appeared -*nsubj*-> piece], [implications -*nmod*-> Qaeda], [passing -*case*-> of] and [Arafat -*case*-> 's]. In this position, we find a more

complicated center-embedded construction: [piece on the [implications [of [Arafat 's] passing] for Qaeda] appeared]. The flux weight is the size of the largest disjoint sub-flux. For position 1 the weight is 1, 2 for position 2, 4 for position 3.

2.2. Granularity

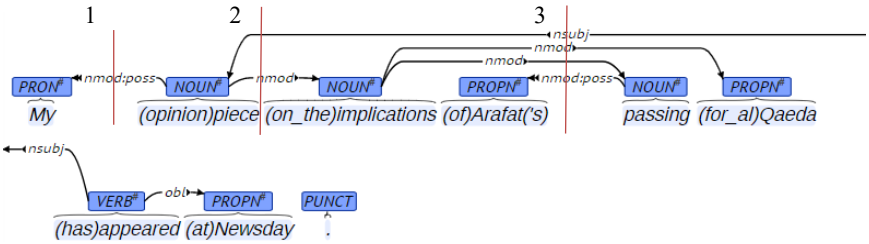

Figure 2. Aggregated tree, tokens in parenthesis are ignored.

At the modeling level, functional relations do not have a unified behavior in every treebank, and some of them are language specific relations. For example, the relation *clf* (classifier) exists in only a few languages, such as Chinese, and can form an additional disjoint dependency in comparison with other languages. It is possible to adjust the granularity of the syntactic analysis in order to make the different treebanks more comparable, for example by keeping only the content words and eliminating relations of the kind: auxiliary, case, conjunction, non-personal relations such as expletives, determiners, and parataxis.

This gives us an aggregated tree, such as the one in Figure 2. The three positions considered in Figure 1 are still marked by a vertical line. In position 1, there remains only one relation in the flux. In position 2, there are no longer any disjoint dependencies. The flux weight in position 1 and position 2 is now 1. For position 3, we have 3 disjoint dependencies, [appeared *-nsubj->* piece], [implications *-nmod->* Qaeda] and [passing *-nmod :poss->* Arafat] and the flux weight is then 3.

3. Dependency flux combined with dependency length

Our hypothesis of the complexity for sentence processing considers two aspects. On the one hand, the complexity depends on the number of disjoint dependencies that we measure by flux weight; on the other hand, it depends on the dependency length (modulo granularity). Thus, by combining the length of dependencies and the flux weight, we introduce a new measure, which we call the *combined weight* of the flux. We would like to look at how the combined

weights behave among the treebanks, as well as to study the characteristics of this new measure.

The combined weight in a given position is the sum of the dependency length of the longest disjoint dependencies. In the aggregated tree of Figure 2, for position 1, we have only one dependency, the length of which is 1, so the combined weight is W_c=1; for position 2, W_c=5; for position 3, W_c= 9 = 5 [appeared-*nsubj*->piece] + 3 [implications-*nmod*->Qaeda] + 1 [passing-*nmod:poss*->Arafat].

3. Results and discussion

3.1. Granularity

By calculating the flux weight of all inter-word positions for every treebank, we found that the maximum weight varies between 3 and 5 in the model of aggregated trees. It is 3 for Vietnamese and Slovak, 10 languages have a maximum weight of 5 and the other 25 treebanks have a flux weight of 4. In comparison with the original treebanks, where the maximum weight varies between 4 (4 treebanks) and 6 (12 treebanks), our model of aggregated trees brings the maximum weight of different treebanks closer. We also obtain the same result for the average weight.

3.2. Combined weight

As shown in Figure 3, the average weight of aggregated trees (AT) is stable and is more universal, because it only considers the center-embedding levels. The average combined weight of aggregated trees (AT) shows slightly the same trend, but it accentuates the differences among the treebanks. The combined weight carries more information about syntactic complexity.

The dependency length takes linear information into account, is correlated with sentence length, and is sensitive to genre (Jiang & Liu, 2015). As we lack information about genre, we cannot determine whether this difference in combined weight is due to different types of languages or to different genres.

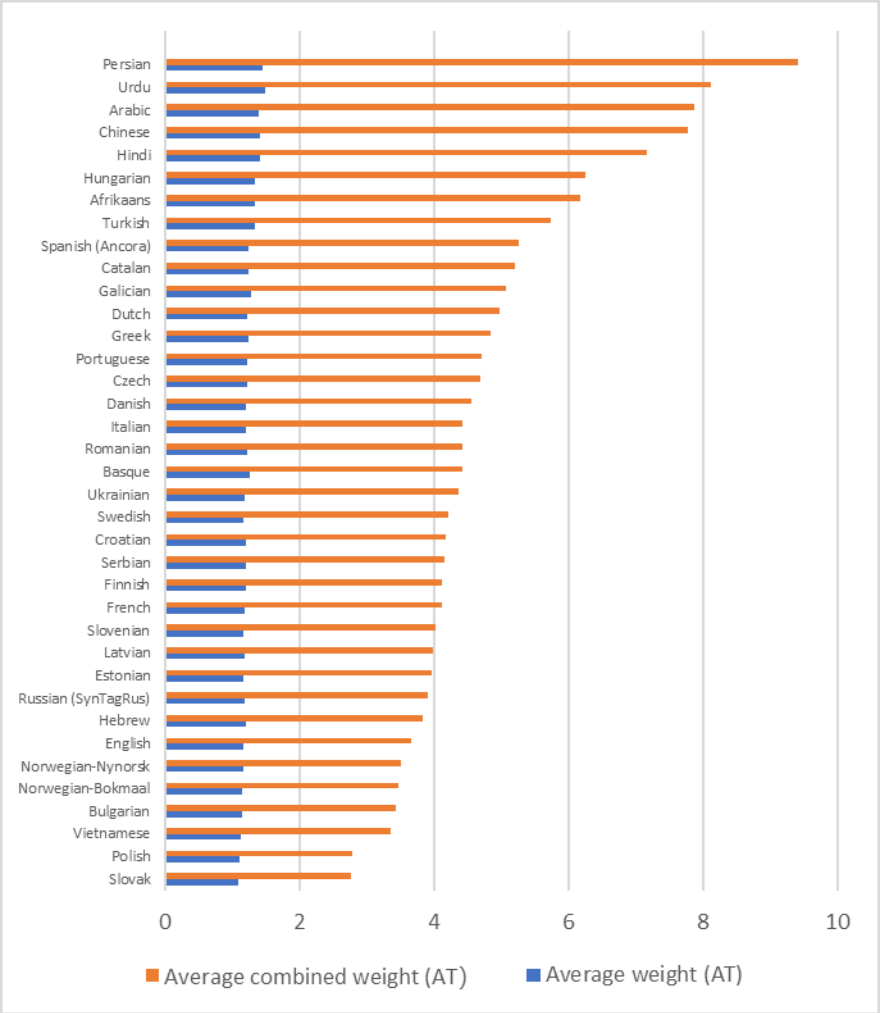

Figure 3. Average combined weight of aggregated trees (AT) and average weight of aggregated trees (AT) in 37 languages. (For more information see the supplementary materials.)

References

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33), 10336-10341.

Jiang, J., & Liu, H. (2015). The effects of sentence length on dependency distance, dependency direction and the implications–Based on a parallel English–Chinese dependency treebank. *Language Sciences*, *50*, 93-104.

Kahane, S., Yan, C., & Botalla, M. A. (2017). What are the limitations on the flux of syntactic dependencies? Evidence from UD treebanks. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)* (pp. 73-82).

Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research*, *25*(1), 93-115.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, *9*(2), 159-191.

Miller, G. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. Psychological Review 63. 81–97.

Miller G. A, Chomsky, N. (1963). Finitary models of language users.

Tesnière, L. (1959). Eléments de la syntaxe structurale. Paris: Klincksieck.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, *104*(5), 444-466.