Proceedings of the Workshop on *Computational Linguistics for Linguistic Complexity* (CL4LC) at the *26th International Conference on Computational Linguistics* (COLING 2016), Osaka, Japan.

A Comparison Between Morphological Complexity Measures: Typological Data vs. Language Corpora

Christian Bentz
University of Tübingen
DFG Center for Advanced Studies
Rümelinstraße 23
chris@christianbentz.de

Tatyana Ruzsics
University of Zürich
CorpusLab, URPP Language and Space,
Freiestrasse 16
tatyana.soldatova@uzh.ch

Alexander Koplenig
Instiute for German Language (IDS)
Mannheim, Germany
koplenig@ids-mannheim.de

Tanja Samardžić
University of Zürich
CorpusLab, URPP Language and Space,
Freiestrasse 16
tanja.samardzic@uzh.ch

Abstract

Language complexity is an intriguing phenomenon argued to play an important role in both language learning and processing. The need to compare languages with regard to their complexity resulted in a multitude of approaches and methods, ranging from accounts targeting specific structural features to global quantification of variation more generally. In this paper, we investigate the degree to which morphological complexity measures are mutually correlated in a sample of more than 500 languages of 101 language families. We use human expert judgements from the *World Atlas of Language Structures* (WALS), and compare them to four quantitative measures automatically calculated from language corpora. These consist of three previously defined corpus-derived measures, which are all monolingual, and one new measure based on automatic word-alignment across pairs of languages. We find strong correlations between all the measures, illustrating that both expert judgements and automated approaches converge to similar complexity ratings, and can be used interchangeably.

1 Introduction

Languages are often compared with regard to their complexity from a computational, theoretical and learning perspective. In computational linguistics, it is generally known that methods mainly developed for the English language do not necessarily transfer well to other languages. The cross-linguistic variation in the amount of information encoded at the level of a word is, for instance, recognized as one of the main challenges for multilingual syntactic parsing (formulated as *The Architectural Challenge* (Tsarfaty et al., 2013)). Complexity of this kind is also found to influence machine translation: translating from morphologically rich languages into English is easier than the other way around (Koehn, 2005). From the perspective of human learning, interesting relationships have been established between the size of populations and morphological complexity (Lupyan and Dale, 2010), as well as the proportion of second language learners and the complexity of case systems: languages with more non-native learners tend to have fewer cases (Bentz and Winter, 2013). These findings are attributed to learning pressures reducing complexity.

An important problem for comparing language complexity is the lack of a standard complexity measure applicable to a wide range of languages and research questions. Many definitions and measures

have been proposed to assess linguistic and, in particular, morphological complexity (Baerman et al., 2015; Sampson et al., 2009). The respective approaches to calculating complexity, and their scope of application, can vary considerably. While factors that need to be taken into account in assessing language complexity are rather well studied, little is known on how different measures relate to each other.

The goal of this paper is to assess the degree to which different language complexity measures are mutually correlated. We are especially interested in the relation between a measure derived from human expert judgements, and several corpus-based measures that do not involve human judgements. We quantify the relations between the measures using 519 languages of overall 101 families represented both in the *World Atlas of Language Structures* (WALS) (Dryer and Haspelmath, 2013) and in parallel corpora (Koehn, 2005; Mayer and Cysouw, 2014). Our findings suggest that the correlation between the measures is strong enough to allow their interchangeable use.

2 Related work

There is a recent rise of interest in defining and measuring linguistic complexity, reflected in three volumes on the topic (Sampson et al., 2009; Baerman et al., 2015; Miestamo et al., 2008). In this spirit, some of our data sources have already been used for quantitative comparisons of a wide range of languages. Lupyan and Dale (2010), for instance, extract indicators of morphological complexity from the WALS, and relate them to the size of speaker populations. In a similar vein, Bentz and Winter (2013) and Bentz et al. (2015) investigate the relationship between morphological complexity, lexical diversity, and the proportion of adult second language learners in speaker populations. We adopt an approach to quantifying the typological data similar to the one presented in these studies, but we adapt it to the needs of our comparison.

The idea of using parallel corpora for language complexity comparison dates back to Greenberg (1959). It was revived with the development of large parallel corpora and computational tools for their processing. Cysouw and Waelchli (2007) provide an overview of how massively parallel texts can be applied to cross-linguistic studies, and describe the potential of such corpora. Recently, massively parallel corpora have been used for studies in lexical typology (Waelchli and Cysouw, 2012), and word order typology (Östling, 2015).

Finally, a new data set specifically intended for information content comparison is under construction at Google (Sproat et al., 2014). The goal of this project is to provide maximally parallel sentences in a set of languages with detailed functional glosses. Once completed, this data set will enable more comprehensive complexity measures than those used in our paper, however, for a relatively small set of languages.

3 Measures

In this section, an overview of the measures used is given. We start with the quantification of expert judgements extracted from the WALS. Next, we move to four corpus-based approaches using type-token ratios, unigram word entropy, relative entropy of word structure, and word alignments.

3.1 Typological measure based on WALS: C_{WALS}

We choose 28 chapters/features of the *World Atlas of Language Structures* (Dryer and Haspelmath, 2013) which are relevant for describing morphology. For example, Chapter 30A "Number of Genders" gives a range of 5 values from "None" to "5 or more", which we directly map to values 1 to 5 indicating increasing complexity. Some features are binary. For instance, Chapter 67A on "The Future Tense", gives a binary distinction between whether there is a morphological marker or not. We code this as 0 and 1. In other chapters such as 70A "The Morphological Imperative" the values have to be reordered to reflect an increasing complexity of morphology. Details about the chapters, their categories, the necessary transformations, and the final values are given in Appendix 8.1.

We arrive at 28 WALS features of morphology with values ordered by increasing use of morphology to encode the feature. There are 1713 languages in WALS for which at least 1 feature value is given. There are only 10 languages for which all 28 features are available. Note, however, that our transformations

result in scales of different sizes for different features. To make the values comparable, we normalize all the values to the interval [0,1]. As a normalization factor we use each feature's maximum value, so that the value of 1 across all the features corresponds to the maximum use of morphology.

Based on the obtained data set we assign a morphological complexity score to each language by averaging the values of the features:

$$C_{WALS} = \frac{\sum_{i=1}^{n} f_i}{n}, \tag{1}$$

where f_i is the feature value of feature i, and n is the number of features available per language. Hence, C_{WALS} is the feature value average per language.

Table 3.1 gives C_{WALS} values for the subset of 34 languages which are represented by either 27 or 28 features.

ISO	Name	Family	No. Chapters	C_WALS
tur	Turkish	Altaic	27	0.775
evn	Evenki	Altaic	27	0.748
abk	Abkhaz	Northwest Caucasian	28	0.704
zul	Zulu	Niger-Congo	27	0.684
swh	Swahili	Niger-Congo	27	0.675
qvi	Quechua (Imbabura)	Quechuan	28	0.662
eus	Basque	Basque	28	0.647
apu	Apurina	Arawakan	27	0.573
lez	Lezgian	Nakh-Daghestanian	28	0.568
arz	Arabic (Egyptian)	Afro-Asiatic	28	0.563
hun	Hungarian	Uralic	28	0.558
heb	Hebrew (Modern)	Afro-Asiatic	27	0.529
wyb	Ngiyambaa	Pama-Nyungan	27	0.528
ckt	Chukchi	Chukotko-Kamchatkan	28	0.519
khk	Khalkha	Altaic	27	0.516
tiw	Tiwi	Tiwian	27	0.495
hix	Hixkaryana	Cariban	27	0.489
hae	Oromo (Harar)	Afro-Asiatic	27	0.487
jpn	Japanese	Japanese	27	0.474
aey	Amele	Trans-New Guinea	27	0.456
rus	Russian	Indo-European	28	0.453
ell	Greek (Modern)	Indo-European	28	0.452
spa	Spanish	Indo-European	27	0.440
deu	German	Indo-European	27	0.397
kut	Kutenai	Kutenai	28	0.357
ind	Indonesian	Austronesian	28	0.336
eng	English	Indo-European	28	0.329
hau	Hausa	Afro-Asiatic	28	0.322
plt	Malagasy	Austronesian	27	0.309
ayz	Maybrat	West Papuan	27	0.292
rap	Rapanui	Austronesian	27	0.218
mri	Maori	Austronesian	27	0.194
yor	Yoruba	Niger-Congo	28	0.178
vie	Vietnamese	Austro-Asiatic	27	0.141

Table 1: Morphological complexity values according to features represented in the WALS. This is a subset of 34 languages with 27 or 28 feature values, though not necessarily of the same features.

3.2 Corpus-based measures

Word entropy: C_H We also measure morphological complexity using *word entropy* (C_H) as described in Bentz and Alikaniotis (2016). This reflects the *average information content* of words. By trend, languages that have a wider range of word types, i.e. packing more information into word structure, rather than phrase or sentence structure, will score higher on this measure.

A "word" is here defined as a unigram, i.e. a string of alpha-numeric Unicode characters delimited by white spaces. Let T be a text that is drawn from a vocabulary of word types $\mathcal{V} = \{w_1, w_2, ..., w_V\}$ of size $V = |\mathcal{V}|$. Further assume that word type probabilities are distributed according to $p(w) = Pr(T = w)$ for $w \in \mathcal{V}$. The average information content of word types can then be calculated as (Shannon and Weaver, 1949)

$$H(T) = -\sum_{i=1}^{V} p(w_i) \log_2(p(w_i)). \tag{2}$$

A crucial step to estimate $H(T)$ is to get word type probabilities $p(w_i)$. The *maximum likelihood* or *plug-in* estimator just takes type frequencies normalized by the overall number of tokens. However, this estimator underestimates the entropy, as it does not take into account unseen types, which is especially

problematic for small texts (Hausser and Strimmer, 2009). A method with a faster convergence rate is the *James-Stein shrinkage* estimator (Hausser and Strimmer, 2009). Word probabilities are here estimated as

$$\hat{p_{w_i}}^{shrink} = \lambda \hat{p_{w_i}}^{target} + (1 - \lambda)\hat{p_{w_i}}^{ML}, \tag{3}$$

where $\hat{p_i}^{ML}$ denotes the word probability according to the maximum likelihood account, $\lambda \in [0, 1]$ is the "shrinkage intensity", and $\hat{p}_i^{target}$ is the "shrinkage target", namely the maximum entropy case of a uniform $p_{w_i} = \frac{1}{V}$. Hausser and Strimmer (2009) illustrate that the optimal shrinkage parameter λ can be found analytically. Given this parameter, the probability $\hat{p}_{w_i}^{shrink}$ plugged into the original entropy equation yields

$$\hat{H(T)}^{shrink} = -\sum_{i=1}^{r} \hat{p_{w_i}}^{shrink} \log_2(\hat{p_{w_i}}^{shrink}). \tag{4}$$

Relative entropy of word structure: C_D C_D is taken from Koplenig et al. (2016), and inspired by earlier accounts to measure different dimensions of language complexity by making use of Lempel-Ziv compression algorithms (Juola, 1998; Juola, 2008; Montemurro and Zanette, 2011; Ehret and Szmrecsanyi, 2016).[1] Let T be a text that is drawn from an alphabet of characters (not words as above) $\mathcal{A} = \{c_1, c_2, ..., c_A\}$ of size $A = |\mathcal{A}|$. Kontoyiannis et al. (1998) illustrate that the per character entropy of T can then be estimated as

$$\hat{H}(T) = \left[\frac{1}{n} \sum_{i=1}^{n} \frac{l_i}{\log_2(i+1)} \right]^{-1}, \tag{5}$$

where n is the overall number of characters in the text T, and l_i is the length of the longest substring from position i onward that has not appeared before, i.e. in T_1^{i-1}. Note that the average match length l_i is related to the redundancy and predictability in T. If match-lengths are generally long, then there is more redundancy and more predictability, if they are short, then there is less redundancy and less predictability in the text.

To estimate the amount of redundancy/predictability contributed by within-word structure, Koplenig et al. (2016) replace each word token in T by a token of the same length but with characters randomly drawn with equal probability from the alphabet $\mathcal{A}$. The entropy of the original text is then subtracted from the masked text to yield

$$\hat{D} = \hat{H}(T^{masked}) - \hat{H}(T^{original}). \tag{6}$$

The bigger $\hat{D}$, the more information is stored within words, i.e. in morphological regularities. This measure of morphological complexity is denoted C_D in the following.

Type/Token ratios: C_{TTR} We take the ratio of word types over word tokens as a simple baseline measure (Kettunen, 2014). The range of word types is expanded by productive morphological markers. Hence, higher values of C_{TTR} correspond to higher morphological complexity. Given a text T drawn from a vocabulary of word types $\mathcal{V} = \{w_1, w_2, ..., w_V\}$ of size $V = |\mathcal{V}|$ the measure is

$$C_{TTR} = \frac{V}{\sum_{i=1}^{V} fr_i}, \tag{7}$$

where V is the number of types, and fr_i is the token frequency of the i^{th} type.

[1] Note that Koplenig et al. (2016) do not call this a "complexity" metric, since they remain neutral about whether word internal structure is more or less difficult to grasp from the perspective of a learner.

Word alignment based measure: C_A Finally, we consider a measure based on word alignment, which, to our knowledge, has not been implemented before. Word alignment is an essential step in phrase-based statistical machine translation (Koehn et al., 2003). The intuition behind the alignment based approach is that words in morphologically richer languages tend to be translated, and therefore aligned, to several words in a morphologically poorer language. As in the case of C_H and C_{TTR} measures the term "word" is understood here in an orthographic sense.

Word alignment from a source to a target language can result in three different scenarios[2]: a single word in the source language is aligned to a single word ("OneToOne") or several words ("OneToMany") in the target language, or several words in the target language are aligned to a single word in the source language ("ManyToOne"). We illustrate these cases by an example of alignments from English to Russian in Figure 3.2:

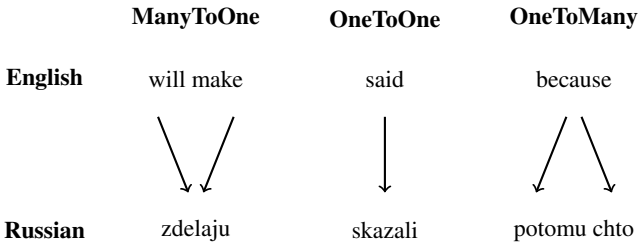

Figure 1: Example of three word alignment categories in a word-aligned English to Russian text.

When word alignments are performed from a morphologically poorer to a morphologically richer language, "ManyToOne" alignments tend to be more frequent than 'OneToMany" alignments, and the other way around. This observation can be quantified using C_A measure:

$$C_A = \frac{\#\mathsf{ManyToOne} - \#\mathsf{OneToMany}}{\#\mathsf{AllAlignments}}, \tag{8}$$

where $\#\mathsf{ManyToOne}$ is the number of all alignments from "ManyToOne" category, $\#\mathsf{OneToMany}$ is the number of all alignments from "ManyToOne" category, $\#\mathsf{AllAlignments}$ is the number of all alignments. A single alignment can be represented by one arrow as in Figure 3.2.

A positive value of C_A indicates that the target language is packaging more information into single words than the source language, negative values correspond to the opposite case. Hence, languages can be compared using C_A values based on word alignments from a fixed source language. An inherent property of the C_A measure is that it is derived based on fully parallel bilingual texts and therefore takes into account direct realisations of how languages encode information through word alignments. The alignment measure is therefore conceptually different from the other three monolingual measures.

4 Data and methods

The corpus-based measures C_D, C_H, C_A and C_{TTR} are calculated using parallel corpora. C_D (Koplenig et al., 2016) is estimated based on the the Book of Matthew (New Testament) from the Parallel Bible Corpus (PBC) (Mayer and Cysouw, 2014). This gives a sample of 1124 so-called "doculects", i.e. indirect representations of languages (defined by ISO-639-3 codes). C_H is estimated based on a sample of parallel texts from the PBC (all books), the *Universal Declaration of Human Rights* (UDHR),[3] and the *European Parliament Corpus* (EPC) (Koehn, 2005). This amounts to 1242 doculects. C_{TTR} is calculated for 1144 doculects of the full PBC.[4]

Since the implementation of C_A requires sentence aligned bitexts with a fixed source language, it is estimated based on PBC with Hebrew being fixed as a source language. However, the complete

[2]Here we assume that the symmetrization heuristic (Och and Ney, 2003) is applied to alignments.

[3]http://unicode.org/udhr/

[4]Note that differences in numbers of doculects can derive from a specific book (e.g. Book of Matthew) not being translated into specific languages.

Bible only exists for around 16% of the languages covered by the PBC. In order to ensure that we use fully parallel texts to calculate the C_A measure, and that these are consistent in terms of the size and content across languages, we only use the New Testament (NT). The average size of the NT appears to be sufficient for producing stable ranking results as confirmed by convergence tests with an increasing amount of parallel verses (see Appendix 8.2).

The PBC as well as the other parallel corpora available for large scale comparative studies is usually rather short in comparison to bilingual text data used to train classical alignment models for machine translation. Therefore we use the Efmaral alignment method (Östling and Tiedemann, 2016) which proves to be the optimal solution for relatively short texts in terms of accuracy and efficiency.

In order to compare different corpus-based measures and the WALS-based measure, we merge the data sets by ISO-639-3 codes. Thus we end up with a data set of 519 languages for which all measures are available.

5 Results

In this section we first investigate how all considered measures of morphological complexity agree between each other using the full dataset and the subsets corresponding to the three biggest language families in our data: Atlantic-Congo, Austronesian, Indo-European, as well as the rest of the 101 families referred further as "Other". Then we proceed with a comparison of how corpus-based measures correlate with the WALS-based measure when we consider increasing subsets of typological features.

5.1 Pairwise correlations between complexity measures

As a result of applying each complexity measure to the data we get a ranked list of languages. Therefore we choose the non-parametric Spearman rank correlation to evaluate associations between each pair of measures.

Figure 2 gives an overview of pairwise correlations between all 5 morphological complexity measures for the full dataset of 519 languages. The density plots on the diagonal panels show the distribution of values for each measure. The lower off-diagonal panels illustrate the correlations between the measures using scatterplots. Each plot shows linear regressions fitted based on subcategorized data: Atlantic-Congo (red), Austronesian (green), Indo-European (blue) and "Other" (purple). The upper panels quantify the correlations for the overall data set (black), and by family (respective colour). Notice that the scale on the y-axis does not apply to the density plots, only to the scatterplots.

The lower off-diagonal scatterplots show that there is always a positive correlation between the different measures, and that this holds across all the four subsets of our data. For the full data set, the correlations between corpus-based measures are generally stronger (ranging from 0.756 to 0.918) than the correlations with C_{WALS} (ranging from 0.318 to 0.437). The strongest correlation is found between C_D and C_{TTR} (0.918), and the weakest between C_{WALS} and C_A (0.318). All correlations reported here are significant at the $p < 0.001$ level.[5]

5.2 Correlations with the WALS measure

Figure 3 focuses on correlations between the typological measure C_{WALS} and the corpus-based measures C_D, C_H,C_A, and C_{TTR}. The values on the left-hand side of the graph correspond to Spearman correlations for each measure calculated over all languages with *at least 1* feature, a total of 519. These numbers correspond to the ones given in the first row of Figure 2. The right-hand side of the graph shows correlations over languages which are represented by *at least 27* features in the WALS, a total of 23.[6]

Note that as we increase the minimum number of features from WALS to be included, we reduce the number of languages (indicated by the size of the dots). As can be seen from the graph, inclusion of more features results in stronger agreement between the corpus-based measures and C_{WALS}. Towards the right hand side we observe the highest values of correlations between C_{WALS} and the corpus-based measures, reaching 0.89 for C_D, 0.88 for C_{TTR}, 0.86 for C_H, and 0.70 for C_A. The intermediate cases

[5]Though see Koplenig (forthcoming) for an argument against hypothesis testing in corpus linguistics.

[6]This is less than the 34 languages in Table3.1 since not all of them are found in the parallel corpora.

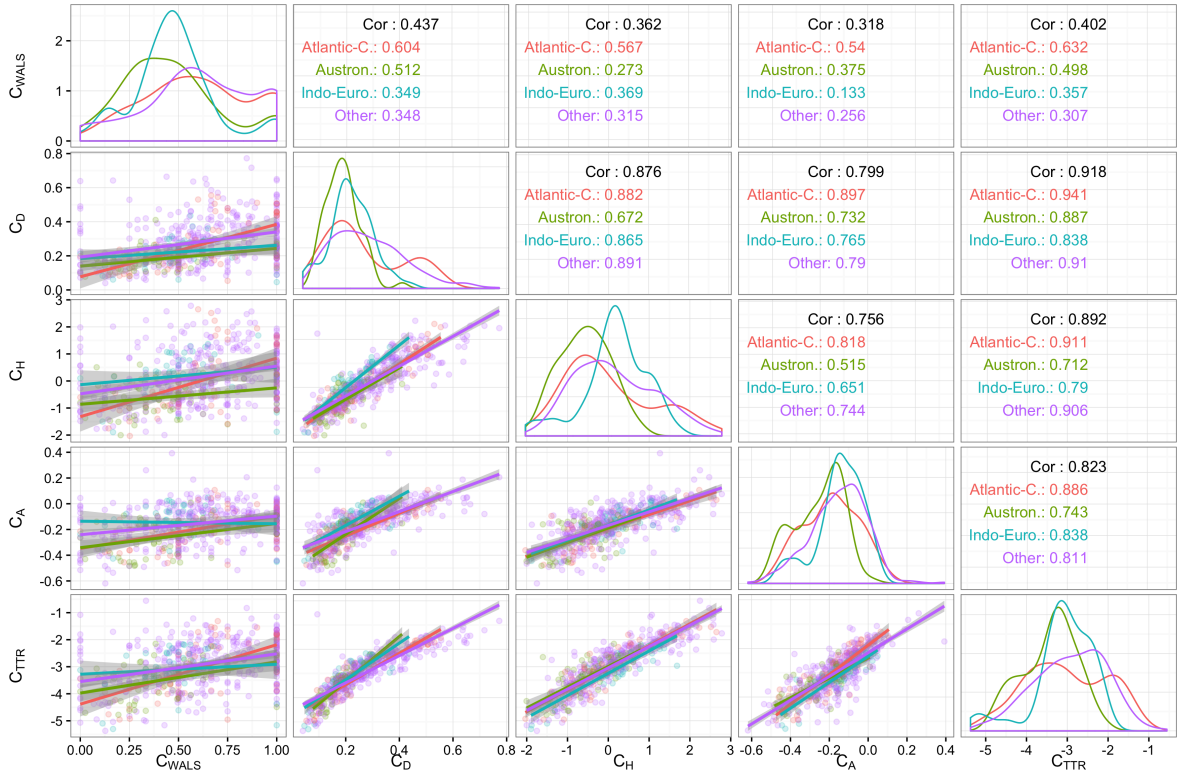

Figure 2: Pairwise correlations between all 5 complexity measures. The lower off-diagonal panels show scatterplots with fitted linear regression lines and 95% confidence intervals. The diagonal panels show density plots for the respective data set. The upper off-diagonal panels give pairwise Spearman correlations. Colours indicate the three major language families Atlantic-Congo (red), Austronesian (green), Indo-European (blue), and the rest subsumed under "Other" (purple). The correlation given in black is the overall correlation. The full data set with 519 languages is used.

have lower coverage of languages as well as less overlap of features between them. This leads to higher disagreement between the measures while they still show common trends. The results illustrate that given enough information in the WALS, C_{WALS} correlates just as good with corpus-based measures as these do amongst each other.

6 Discussion

Our results of comparing different morphological measures provide two major insights:

1. We used four vastly differing automated approaches of measuring morphological complexity (C_D,C_H,C_A, and C_{TTR}) in actual language production, i.e. parallel corpora. They all display strong correlations between each other, i.e. strong agreement on which languages are morphologically complex, and which are not. This is encouraging, since it illustrates that the judgements of these automated methods converge despite the conceptual differences.

2. Given enough feature values, the expert judgements of the WALS also converge with the automated corpus-based methods when ranking languages on a morphological complexity scale, which is reflected in Spearman correlations of up to 0.89. This is remarkable when considering how much expert knowledge and working hours go into writing of descriptive grammars, and assembling them in databases like WALS. If our sole objective is to rank languages in terms of morphological complexity, the automated methods yield high agreement with the outcome of a human expert rating.

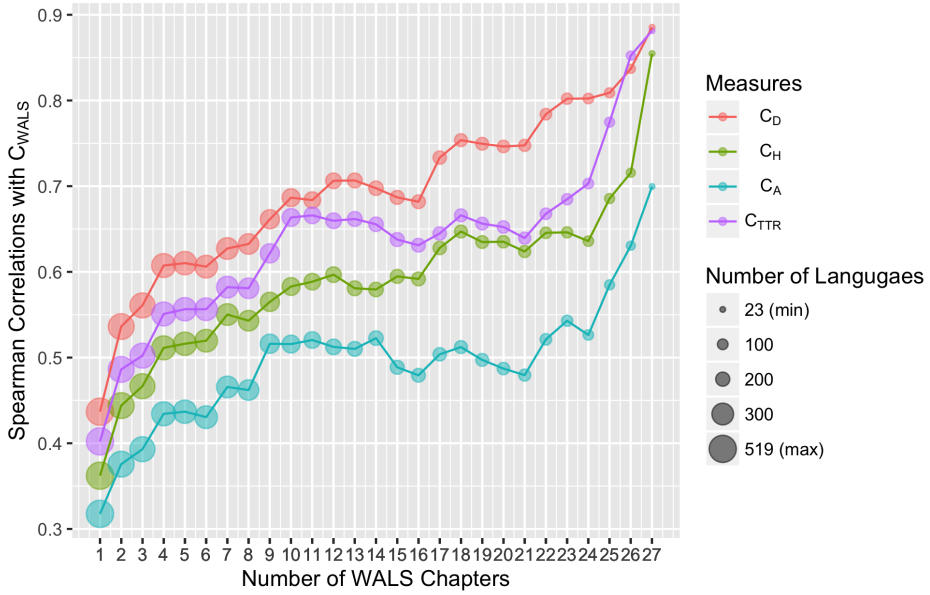

Figure 3: Spearman correlations of the corpus-based measures with C_{WALS}. The x-axis corresponds to a subcategorization of the languages included to calculate the correlations. For example, "27" means that only languages which have values in at least 27 WALS chapters are included, which is the case for 23 languages, and so on. All correlations are significant ($p < 0.001$).

Since all the measures correlate strongly, the reason for choosing either of them depends on the objective and the data limitations of a given study. In the following, we discuss some of the advantages and disadvantages of the respective measures.

6.1 Advantages and disadvantages of the measures

The objective of the WALS is to give an overview of phonetic, morphological, lexical and syntactic properties of a large and balanced sample of languages. It is a collaborative effort of dozens of experts to establish a data base that allows cross-linguistic comparisons. As such, it is a rich source for typological studies. However, by necessity it is only a coarse-grained reflection of the actual dimensions of complexity. For example, classifying a language by whether it uses 2, 3, 4, etc. nominal case markers (chapter 49) does not tell us how productive these markers are in actual language production. Note, also, that the coverage of WALS in terms of features per language is sparse. If we want to include all 28 features, then we end up with a sample of only 10 languages.

The corpus-based measures allow us to look at real instances of morphological productivity in texts across many languages. All of them can be estimated from corpora directly and efficiently, without much prior processing. The only requirement for C_H, C_D and C_{TTR} is that word types are delimited in a consistent manner, e.g. by white spaces – or other non-alphanumeric characters. Hence, these measures come without much theoretical "baggage", and are cross-linguistically comparable. At this point, they can be applied to ca. 1500 languages via massively parallel corpora like the PBC.

The simplest corpus-based measure is C_{TTR}. Once word types are defined, it is easily and straightforwardly computed from a text. However, a drawback of C_{TTR} is that it does not take into account subtle differences in the distributions of word tokens over word types. C_H is a more accurate reflection of the actual distributions. However, just like C_{TTR}, C_H does not distinguish between effects due to breadth of the base lexicon, on one hand, and word formation processes such as derivation, inflection or compounding, on the other. Also, it does not reflect differences in regular and irregular morphological processes. For example, the irregular pair *go→went* will contribute just as much to higher C_H as the regular pair *sprint→sprinted*.

C_D has the advantage of distinguishing regular from irregular processes of word formation. Regular

suffixes as in the example above will introduce systematic redundancy reflected in $\hat{D}$. However, masking within-word structure requires a further processing step that might introduce biases which are not well understood yet. Also, despite picking up on regular patterns within words, C_D does still not distinguish between different types of word formation.

Likewise, C_A does – at this stage – not distinguish between different types of word formation. This could be overcome by considering alignments on the type level. Given the sum of all alignments for a word type, it is expected that the diversity of alignments will be lower for a word in a language with rich morphology, since the word types are expected to be rarer (less frequent). Therefore, a measure based on word type level alignments could distinguish between morphological and lexical diversity. Also, the influence of the choice of source language on the results needs to be clarified in future studies.

6.2 Conceptualizing morphological complexity

"Linguistic complexity" more generally, and "morphological complexity" in particular, are polysemous concepts. Here, we focused on defining and comparing different quantitative measures. They necessarily hinge upon different conceptualizations of complexity. The account based on WALS chapters is a *paradigm-based* approach. Harnessing descriptive grammars, typologists attribute a given number of paradigmatic distinctions to languages, which, in turn, reflect their complexity. Measures such as C_{TTR}, C_D and C_H, on the other hand, could be called *distribution-based*. They conceptualize complexity with reference to the distribution of word tokens over word types used in a given language. Finally, the *translation-based* account C_A is not applicable to single languages, but conceptualizes complexity via the problem of translating a concept from one language to another. Clearly, all of these are conceptualizations in their own right, with specific implications for language learning and usage. However, they turn out to be strongly correlated – across the board – since they all reflect different nuances of the same principle: linguistic complexity relates to the fundamental information-theoretic concept of *uncertainty* or *choice* when encoding and decoding a message.

7 Conclusions

We have tested four conceptually different measures of morphological complexity across more than 500 languages of 101 families. The overall results suggest that different corpus-based measures are highly consistent when ranking languages according to morphological complexity. Moreover, measures based on typological expert judgements are also converging onto similar rankings if enough language specific information is given. These findings help to establish a quantitative, empirical, and reproducible account of morphological complexity, and linguistic typology more generally.

Acknowledgements

CB was funded by a grant for international short visits of the Swiss National Science Foundation (SNSF), as well as the German Research Foundation (DFG FOR 2237: Project "Words, Bones, Genes, Tools: Tracking Linguistic, Cultural, and Biological Trajectories of the Human Past"), and the ERC Advanced Grant 324246 EVOLAEMP.

References

Matthew Baerman, Dunstan Brown, Greville G Corbett, et al. 2015. *Understanding and measuring morphological complexity*. Oxford University Press.

Christian Bentz and Dimitrios Alikaniotis. 2016. The word entropy of natural languages. *arXiv preprint arXiv:1606.06996*.

Christian Bentz and Bodo Winter. 2013. Languages with more second language speakers tend to lose nominal case. *Language Dynamics and Change*, 3:1–27.

Christian Bentz, Annemarie Verkerk, Douwe Kiela, Felix Hill, and Paula Buttery. 2015. Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *PLoS ONE*, 10(6):e0128254.

Michael Cysouw and Bernard Waelchli. 2007. Parallel texts: Using translational equivalents in linguistic typology. *STUF - Language Typology and Universals*, 60:95 – 99.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *World Atlas of Language Structures online*. Max Planck Digital Library, Munich.

Katharina Ehret and Benedikt Szmrecsanyi. 2016. An information-theoretic approach to assess linguistic complexity. In Raffaela Baechler and Guido Seiler, editors, *Complexity and Isolation*. de Gruyter, Berlin.

Joseph Harold Greenberg. 1959. A quantitative approach to morphological typology. *International Journal of American Linguistics*, 26:178–194.

Jean Hausser and Korbinian Strimmer. 2009. Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *The Journal of Machine Learning Research*, 10:1469–1484.

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

Patrick Juola. 2008. Assessing linguistic complexity. In Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, editors, *Language complexity: typology, contact, change*, pages 89–108. Amsterdam: John Benjamins.

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

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Ioannis Kontoyiannis, Paul H Algoet, Yu M Suhov, and Abraham J Wyner. 1998. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *Information Theory, IEEE Transactions on*, 44(3):1319–1327.

Alexander Koplenig, Peter Meyer, Sascha Wolfer, and Carolin Mueller-Spitzer. 2016. The statistical tradeoff between word order and word structure: large-scale evidence for the principle of least effort. *arXiv preprint arXiv:1608.03587*.

Alexander Koplenig. forthcoming. Against statistical significance testing in corpus linguistics.

Gary Lupyan and Rick Dale. 2010. Language structure is partly determined by social structure. *PloS ONE*, 5(1):e8559, January.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014*, pages 3158–3163. European Language Resources Association (ELRA).

Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson. 2008. *Language complexity: Typology, contact, change*. John Benjamins Publishing.

Marcelo A Montemurro and Damián H Zanette. 2011. Universal entropy of word ordering across linguistic families. *PLoS One*, 6(5):e19875.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29:19 – 51.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106, October. To appear.

Robert Östling. 2015. Word order typology through multilingual word alignment. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 205 – 211.

Geoffrey Sampson, David Gil, and Peter Trudgill. 2009. *Language complexity as an evolving variable*. Oxford University Press.

Claude E. Shannon and Warren Weaver. 1949. *The mathematical theory of communication*. The University of Illinois Press, Urbana.

Richard Sproat, Bruno Cartoni, HyunJeong Choe, David Huynh, Linne Ha, Ravindran Rajakumar, and Evelyn Wenzel-Grondie. 2014. A database for measuring linguistic information content. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 967–974, Reykjavik, Iceland, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1397.

Reut Tsarfaty, Djam Seddah, Sandra Kbler, and Joakim Nivre. 2013. Parsing morphologically rich languages: Introduction to the special issue. *Computational Linguistics*, 39(1):15–22.

Bernard Waelchli and Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics*, 50:671 – 710.

8 Appendices

8.1 Transformations of WALS features

Chapter	Name	Categories	Transformation	Final Values
22A	Inflectional Synthesis	7 (ordinal)	none	1-7
26A	Prefixing vs. Suffixing in Inflectional Morphology	6 (non-ordinal)	binarization	0-1
27A	Reduplication	3 (non-ordinal)	binarization	0-1
28A	Case Syncretism	4 (ordinal)	reorder	1-4
29A	Syncretism in Verbal Person/Number marking	3 (ordinal)	none	1-3
30A	Number of Genders	5 (ordinal)	none	1-5
33A	Coding of Nominal Plurality	9 (partially ordinal)	binarization	0-1
34A	Occurrence of Nominal Plurality	6 (ordinal)	none	1-6
37A	Definite Articles	5 (non-ordinal)	binarization	0-1
38A	Indefinite Articles	5 (non-ordinal)	binarization	0-1
49A	Number of Cases	9 (ordinal)	remove	1-8
51A	Position of Case Affixes	9 (non-ordinal)	binarization	0-1
57A	Position of Pronominal Possessive Affixes	4 (non-ordinal)	binarization	0-1
59A	Possessive Classification	4 (ordinal)	none	1-4
65A	Perfective/Imperfective Aspect	binary	none	0-1
66A	The Past Tense	4 (ordinal)	reorder	1-4
67A	The Future Tense	binary	none	0-1
69A	Position of Tense/Aspect Affixes	5 (non-ordinal)	binarization	0-1
70A	The Morphological Imperative	5 (partially ordinal)	recategorization	1-4
73A	The Optative	binary	none	0-1
74A	Situational Possibility	3 (non-ordinal)	binarization	0-1
75A	Epistemic Possibility	3 (non-ordinal)	binarization	0-1
78A	Coding of Evidentiality	6 (non-ordinal)	binarization	0-1
94A	Subordination	5 (non-ordinal)	binarization	0-1
101A	Expression of Pronominal Subjects	6 (non-ordinal)	binarization	0-1
102A	Verbal Person Marking	5 (partially ordinal)	recategorization	1-3
111A	Nonperiphrastic Causative Constructions	4 (non-ordinal)	binarization	0-1
112A	Negative Morphemes	6 (non-ordinal)	binarization	0-1

Table 2: Recoding of WALS chapters. The column "Categories" gives the type of of the original WALS variable represented in the respective chapter. "Transformations" is a short description of how the chapters where recoded. "Final Values" gives the range of ordinal values used to reflect the morphological complexity.

8.2 Convergence tests for C_A

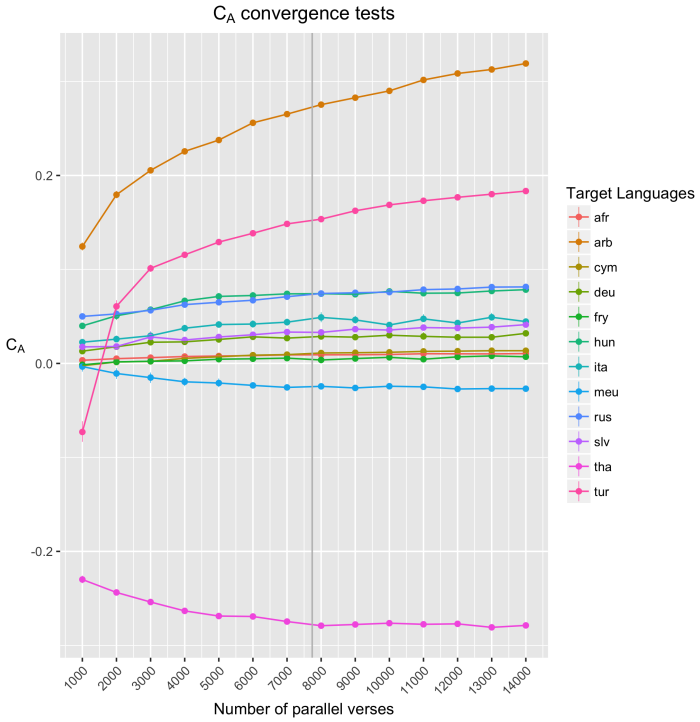

Figure 4: Convergence tests for the C_A measure. The source language for alignments is English. For each size of the verse number X the C_A measure is averaged across 50 samples with X randomly drawn parallel verses. The grey line corresponds to the average number of verses available for NT in PBC.