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

Christian Bentz, University of Tübingen
Tatyana Ruzsics, University of Zürich
Alexander Koplenig, Institute for German Language Mannheim
Tanja Samardžić, University of Zürich
Contact: chris@christianbentz.de

Introduction

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.

1 Measures

Typological

C_WALS: Average complexity value based on up to 28 features on morphology of the World Atlas of Language Structures (Dryer et al. 2013)

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

f_i= value per feature
n=number of features

Corpus-based

C_H: The word entropy as calculated from parallel texts (Mayer & Cysouw, 2014)

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

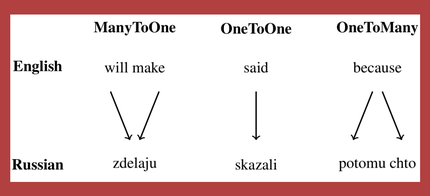
w_i= word type
V= number of word types
T= a given parallel text

C_D: The difference in character entropy before and after word internal regularities have been masked (Koplenig et al., forthcoming)

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

T^{masked}: parallel text with masked regularities
$T^{original}$: original parallel text

C_A: normalized difference in word alignments from a fixed source language to a target language

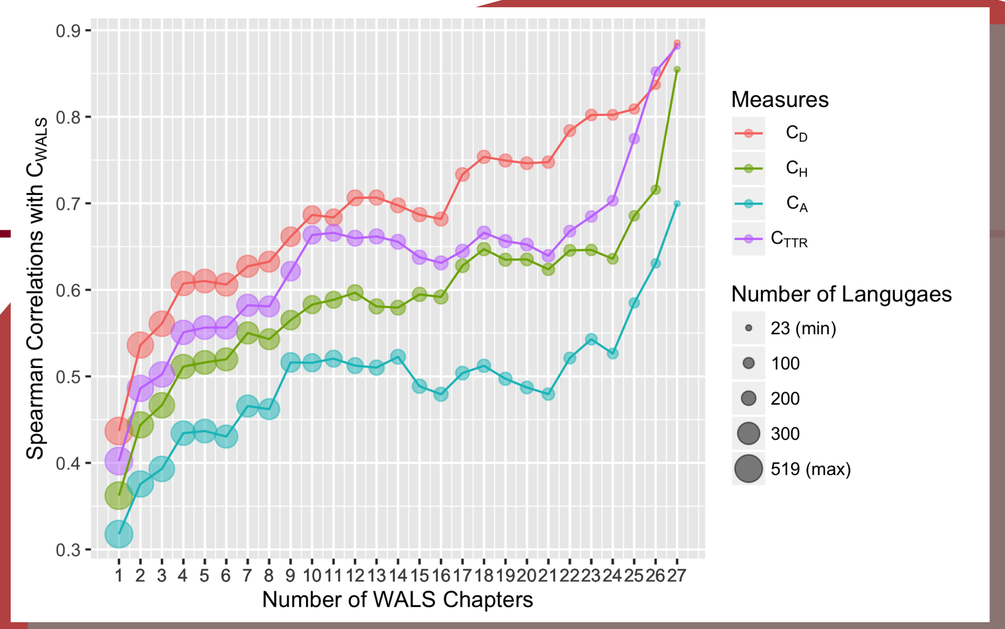


$$C_A = \frac{\#\text{ManyToOne} - \#\text{OneToMany}}{\#\text{AllAlignments}}$$

C_TTR: type-token ratio for parallel texts

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

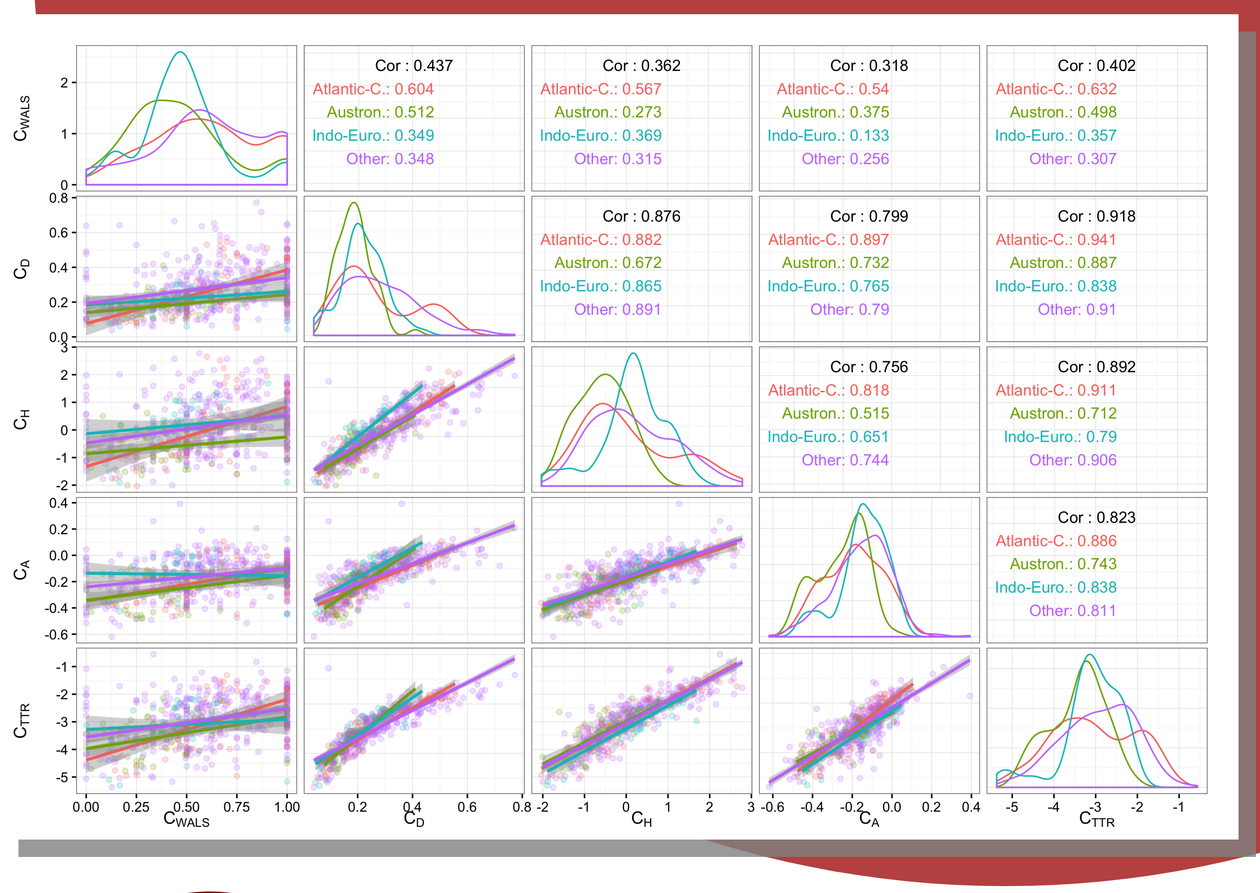
V= number of word types
fr= token frequency of ith word type



2 Results

Depending on the number of WALS features included (1–27), **Spearman correlations** between C_WALS and the corpus-based measures range from **0.3 to 0.9** (figure above)

The correlations hold between different language families (below)



3 Implications

1. All corpus-based automated measures display strong correlations between each other, i.e. **strong agreement on which languages are morphologically complex**. This is the case despite the conceptual differences between automated methods.

2. Given enough feature values, the **expert judgements of the WALS also converge with the automated corpus-based methods**. If our sole objective is to rank languages on a morphological complexity scale, then automated methods can support human expert rating.

References

Matthew S. Dryer & Martin Haspelmath. 2013. World Atlas of Language Structures online. *Max PlanckDigital Library*, Munich.
Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. In Nicoletta Calzolari et al., *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC-2014), Reykjavik, Iceland, May 26-31.
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.