What is *morphologically rich*?

Mark Anderson
Prifysgol Caerdydd/Cardiff University
andersonm8@cardiff.ac.uk

1. Introduction

People often talk of *morphologically rich* languages in NLP but don't really establish what that means. Is Polish with its involved case system morphologically rich or is English with its complicated comparative/superlative system? Or is morphologically rich meant to be left to the purview of agglutinative nightmares where an entire *sentence* can be encoded as a single *word*[1] (Kuriyozov, Doval, and Gómez-Rodríguez 2020)? Here, I give details of the measurements included in the corresponding code[2] which gives an estimation of the morphological complexity that a treebank exhibits. It *doesn't* measure the morphological *richness* of a given language, but it does allow people to compare different treebanks and at least say something like, "Treebank X appears to be more morphologically complex than treebank Y."

2. Side note on WALS

While we could try to utilise actual linguistic information about topological features relating to morphology from resources like the World Atlas of Language Structures (WALS) (Dryer and Haspelmath 2013), it isn't trivial because despite WALS being an incredibly expansive and useful resource, the features pertinent to this discussion only cover a subsection of extant languages. More importantly the coverage for UD treebanks is low. For example, I looked at features 20A (fusion), 21A (case exponence), 21B (TAM exponence), 22A (verb synthesis), 23A (locus marking clause), 24a (locus marking possessives), 27A (reduplication), 28A (case syncretism), 29A (verbal syncretism), 49A (number of cases), and 26A (affixation). In UD v2.5 for the treebanks with a pretrained UDPipe 1.2 model, only 23 out 57 languages have more than 2 features. However, those which have some features typically have them all, i.e. 18 out 57 for UD v2.5.

At a quick glance, it looks like of the 122 languages covered in UD v2.9, 61 are mentioned in at least one of these features, while 49 have 2 or more, and 23 are covered in all the features. This is without checking to see if there are issues with the names of the treebanks in UD and with only a wee bit of jiggery-pokery to try to get as much coverage as possible (e.g. "Scottish Gaelic" to "Gaelic (Scots)"; "Arabic" to "Arabic (Egyptian)"; "Serbian" to "Serbian-Croatian"; "Slovenian" to "Slovene"; "Chinese" to "Mandarin"; "Greek" to "Greek (Modern)"; "Hebrew" to "Hebrew (Modern)"; "Croatian": "Serbian-Croatian"; and "Urdu": "Hindi"). Some changes might be contentious (or straight up erroneous), but I am just wanting to highlight that the coverage of WALS features pertaining to morphology (the set of features which I have used not necessarily being

1 What is a word (Haspelmath 2011)?
2 https://github.com/markda/morphological-complexity.

exhaustive) is patchy to say the least. And raises questions as to how to aggregate a score across incomplete feature vectors. Bentz et al. (2016) do try to use WALS, so you can see some work trying to relate this to parsing difficulty.

3. An aggregate measurement of morphological complexity

I'm just giving the details about the measurement we used for approximating morphologically rich subsets of the treebanks used in one of our experiments. It doesn't give a measurement of how *morphologically rich* a language is, but rather gives a measurement of a subsample of a language found in a given treebank of the characteristics it exhibits that are at least somewhat related to morphology. It isn't anything new or snazzy, but rather bringing a few things together. It is clearly heavily influenced by the work found in Dehouck (2019), specifically chapter 7.

It is an aggregate measurement, consisting of word entropy (Shannon 1948), type-token ratio (Bentz et al. 2016), form to lemma ratio, inflected to form ratio, and head part-of-speech entropy (Dehouck and Denis 2018). These are normalized when needed such that 0 means no morphological complexity and 1 means the highest possible morphological complexity, so that we can simply take the mean measurement across all 5 metrics.

3.1 Normalized word entropy

Word entropy gives an indication as to how much information any given word has with a higher entropy resulting from a treebank having many forms. It is given by:

$$H_{\text{word}} = -\sum_{v \in \mathcal{V}} p(v) \log_2 p(v) \tag{1}$$

where $\mathcal{V}$ is the vocab space in a given treebank, v is a given word in that space, and $p(v)$ is the probability of that word occurring estimated by its frequency count (Shannon 1948). The normalized word entropy, H^*_{word}, is obtained by dividing by the log of the magnitude of the vocab space:

$$H^*_{\text{word}} = \frac{H_{\text{word}}}{log_2|\mathcal{V}|} \tag{2}$$

3.2 Type token ratio

The type-token ratio gives an indication of the morphological production in a given treebank. It is given by:

$$TTR = \frac{|\mathcal{V}|}{|T|} \tag{3}$$

where $\mathcal{V}$ is the vocab space in a given treebank and T is the number of tokens (Bentz et al. 2016). While this number isn't exactly bounded by 0 at the lower margin (it is bounded by 1 at the uppper margin), when T is suitably big, which is typically the case, the instance where $\mathcal{V}$ only consists of 1 type, TTR tends to zero. However, this is clearly not a likely scenario in a treebank and so this inconsistency is not a worry in reality.

3.3 Form to lemma ratio

The form to lemma ratio is similar to the type-token ration but it more closely measures morphological production by honing in on lemmas having multiple forms rather than just looking at the more global measurement of production in TTR. It is given by:

$$F/L = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} |\mathcal{F}_l| \tag{4}$$

where $\mathcal{L}$ is the lemma vocab of a treebank, l is a given lemma in the vocab, and $\mathcal{F}_l$ is the set of forms associated with l (Dehouck and Denis 2018).

As defined F/L ranges from 1 to $|\mathcal{V}|$ (the absurd case of a singular lemma). By taking the reciprocal, we obtain a value that tends to zero in the absurd case and has an upper bound of 1. However, this gives us an inverse scale, i.e. a lower value means more morphology and a higher value less. Therefore we subtract the reciprocal of F/L from 1:

$$F/L^* = 1 - \frac{1}{F/L} \tag{5}$$

3.4 Inflected form to lemma ratio

This is the same as F/L but for the case where a lemma is actually inflected, i.e. the case where the set of word forms associated with a given lemma is greater than 1. It is given by:

$$F/iL = \frac{1}{|\mathcal{L}_2|} \sum_{l \in \mathcal{L}_2} |\mathcal{F}_l| \tag{6}$$

where $\mathcal{L}_2$ is the subset of lemmas which have 2 or more forms associated with them in a treebank, l is a given lemma in that subset, and $\mathcal{F}_l$ is the set of forms associated with l (Dehouck and Denis 2018). It is normalized in the same way as F/L:

$$Fi/L^* = 1 - \frac{1}{F/iL} \tag{7}$$

3.5 Head part-of-speech entropy

The head part-of-speech entropy (HPE) is the most measurement of morphology most related to parsing as it captures the morphosyntactic complexity found in a treebank. It is measured of the delexicalised version of the treebank, where the unit is a concatenation of a token's POS tag and morphological feature tags. The HPE of a treebank is an average over the HPE of each delexicalized word type:

$$HPE = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} HPE_d \tag{8}$$

where:

$$HPE_d = -\sum_{t \in \mathcal{T}_d} p(h_d^t) \log_2 p(t_d^t) \tag{9}$$

where h_d^t denotes the head of d having the POS tag t from the tagset $\mathcal{T}_d$ (the set of tags that d is headed by in the treebank) and $p(h_d^t)$ is the probability of this occuring based on frequency counts (Dehouck and Denis 2018). As defined this gives a value that tends to zero when morphosyntactic complexity is prevalent and increases unbounded the less morphosyntactic complexity is present. In order to normalize this, we have to normalize HPE_d:

$$HPE_d^* = \frac{HPE_d}{\log_2 |\mathcal{T}_d|} \tag{10}$$

such that the normalized head part-of-speech entropy is simply:

$$HPE^* = 1 - \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} HPE_d^* \tag{11}$$

Note that the sum over the normalized HPE_d values is subtracted from to inverse the scale such that 0 denotes no morphosyntactic complexity and 1 the maximum.

3.6 Aggregate metric

The final metric we used is a simple unweighted average of the 5 normalized metrics described above:

$$MC = \frac{(H^*_{\text{word}} + TTR + F/L^* + F/iL^* + HPE^*)}{5} \tag{12}$$

Below are the lists with the treebanks considered *morphologically complex* in UD v2.5 and v2.6, respectively. The corresponding code is available at `https://github.com/markda/morphological-complexity`.

3.7 List of morphologically rich treebanks in UD v2.5

Ancient Greek-PROIEL	Gothic-PROIEL	Persian-Seraji
Ancient Greek-Perseus	Greek-GDT	Polish-LFG
Armenian-ArmTDP	Hungarian-Szeged	Polish-PDB
Basque-BDT	Irish-IDT	Portuguese-GSD
Belarusian-HSE	Latin-ITTB	Romanian-Nonstandard
Bulgarian-BTB	Latin-PROIEL	Romanian-RRT
Croatian-SET	Latin-Perseus	Russian-GSD
Czech-CAC	Latvian-LVTB	Russian-SynTagRus
Czech-CLTT	Lithuanian-ALKSNIS	Russian-Taiga
Czech-FicTree	Lithuanian-HSE	Serbian-SET
Czech-PDT	Maltese-MUDT	Slovak-SNK
Estonian-EDT	Marathi-UFAL	Slovenian-SSJ
Estonian-EWT	North Sami-Giella	Slovenian-SST
Finnish-FTB	Old Church Slavonic-PROIEL	Tamil-TTB
Finnish-TDT	Old French-SRCMF	Telugu-MTG
German-HDT	Old Russian-TOROT	Turkish-IMST

3.8 List of morphologically rich treebanks in UD v2.6

Ancient Greek-PROIEL	Greek-GDT	Old Russian-TOROT
Ancient Greek-Perseus	Hungarian-Szeged	Persian-Seraji
Armenian-ArmTDP	Irish-IDT	Polish-LFG
Basque-BDT	Latin-ITTB	Polish-PDB
Belarusian-HSE	Latin-PROIEL	Portuguese-GSD
Bulgarian-BTB	Latin-Perseus	Romanian-RRT
Croatian-SET	Latvian-LVTB	Russian-GSD
Czech-CAC	Lithuanian-ALKSNIS	Russian-Taiga
Czech-CLTT	Lithuanian-HSE	Sanskrit-Vedic
Czech-FicTree	Maltese-MUDT	Serbian-SET
Estonian-EDT	Marathi-UFAL	Slovak-SNK
Estonian-EWT	North Sami-Giella	Slovenian-SSJ
Finnish-FTB	Old Church Slavonic-PROIEL	Slovenian-SST
Finnish-TDT	Old French-SRCMF	Tamil-TTB
Gothic-PROIEL	Old Russian-RNC	Telugu-MTG

References

Bentz, Christian, Tatjana Soldatova, Alexander Koplenig, and Tanja Samardžić. 2016. A comparison between morphological complexity measures: Typological data vs. language corpora. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 142–153.

Dehouck, Mathieu. 2019. *Multi-Lingual Dependency Parsing: Word Representation and Joint Training for Syntactic Analysis. (Parsing en Dépendances Multilingue: Représentation de Mots et Apprentissage Joint pour l'Analyse Syntaxique)*. Ph.D. thesis, Université de Lille.

Dehouck, Mathieu and Pascal Denis. 2018. A framework for understanding the role of morphology in Universal Dependency parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2864–2870, Association for Computational Linguistics, Brussels, Belgium.

Dryer, Matthew S. and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 45(1):31–80.

Kuriyozov, Elmurod, Yerai Doval, and Carlos Gómez-Rodríguez. 2020. Cross-lingual word embeddings for Turkic languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4054–4062, European Language Resources Association, Marseille, France.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.