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.

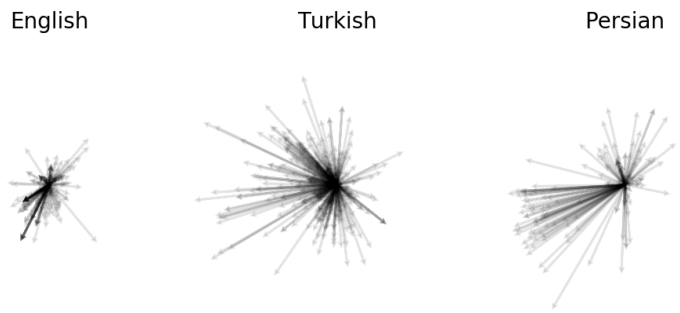

Figure 1. Distance and angle of all difference vectors (traversals between lemma and wordform) in three languages, projected into two-dimensional vector space and arranged around a common origin.

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

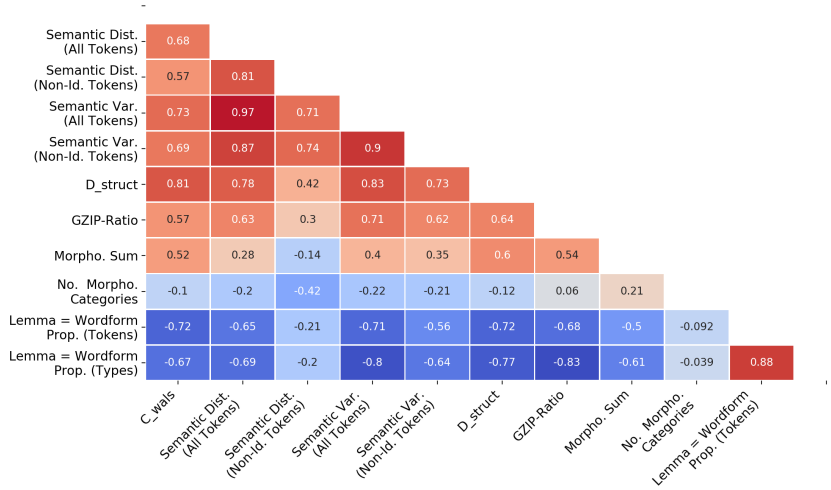

Figure 2. Correlation among our proposed measures, existing measures, and lower level morphological summary statistics.

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 Nynorks 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.*