TAMMI 2.0 is a freely available text natural language processing (NLP) tool specifically designed to annotate and count morphological features in texts. The source code for the tool is available at https://github.com/scrosseye/tammi.[1] The downloadable version of TAMMI 2.0 is available for both Mac and Windows operating systems and features a graphical user interface that allows users to batch process texts.

TAMMI 2.0 provides automatic calculations for the MorphoLex data frame provided by Sánchez-Gutiérrez et al. (2017). MorphoLex was assessed for reliability in a series of analyses that examined the influence of MorphoLex variables on the lexical decision latencies of 4,724 morphologically complex nouns. TAMMI 2.0 also includes an automatic calculation of morphological complexity index (MCI) based on inflections as detailed by Brezina and Pallotti (2019). In addition, TAMMI 2.0 calculates an MCI for derivational morphemes and developed new morphological complexity indices based on morphological variety and type-token ratios for both inflectional and derivational morphemes. Lastly, TAMMI 2.0 calculates a number of basic morpheme counts. All morpheme indices included in TAMMI measure morphological complexity.

Pre-processing. Before calculating any morpheme features, texts are pre-processed using spaCy's small, English model based on the web (Honnibal & Montani, 2017). spaCy is used to first tokenize the texts to extract the words. Next, stop words from spaCy's stop word list are removed. The stop word list includes 326 stop words and contractions. Stop words include pronouns, copula verbs, conjunctions and connectives, prepositions, some adverbs like *mostly, formerly,* and *really,* and delexical verbs like *give, have, done,* and *be*. Contractions include *'ll, 're,* and *n't*. Lastly, only alphabetic tokens are kept. This removes all numbers and punctuation.

[1] This script does not include the graphical user interface code.

Basic morpheme counts. TAMMI 2.0 includes basic morpheme counts for inflections and derivational morphemes. The inflections are counted using spaCy by assessing the differences in the number of characters between each token and its lemma. spaCy uses rule-based approaches for English that include part of speech (POS) tagging to assign base forms (i.e., uninflected forms) to tokens. For instance, *swimming* used as a noun (i.e., gerund) would not be lemmatized but *swimming* as a verb would be lemmatized to *swim*. Similar results would hold for *boring* used as a participial adjective, which would not be lemmatized, and *boring* or *bored* used as a verb, which would both be lemmatized to *bore*. The precision of lemmatization depends on the spaCy POS tagger, which reports an accuracy of ~98%[2] for the larger models when evaluated against the OntoNotes 5.0 corpus (Weischedel et al., 2013); however, this accuracy is likely lower for smaller models and for data on which the tagger was not trained.

Derivational morphemes are calculated using MorphoLex, which provides counts for prefixes, suffixes, and affixes as well as the number of compound words (i.e., words that have more than one root morpheme). In addition, TAMMI calculates the average number of morphemes per word. All basic morpheme counts are normed by the number of content words in the text (i.e., all words that are not in the spaCy stop word list). TAMMI 2.0 also computes normed indices by taking the count for each variable and dividing it by the number of content words with the relevant morpheme. For example, when calculating the number of suffixes, TAMMI 2.0 will sum the number of suffixes in the text and norm that sum by the number of words that contain a suffix. The same is done for affixes and total suffixes.[3] Features normed by relevant morpheme type will likely not perform well on texts with simple morpheme use because each word may only contain a single morpheme type (i.e., a text may contain only a single suffix

[2] Reliability for spaCy is reported at https://spacy.io/usage/facts-figures.
[3] A worked example for calculating Morpholex scores based on a single sentence is available at https://github.com/scrosseye/Tammi-Analyses

in all words that contain a suffix, giving a normed score by suffix of 1). Thus, users should only use the relevant normed morpheme counts when examining longer texts that are representative of more advanced language use.

Morphological variety. The inflection morphological variety feature in TAMMI 2.0 is based on a within-subset variety score in which content words from each text are broken into windows of 10 words (plus a window of 1-to-9 for any remaining content words at the end of the text). Inflectional morpheme types (e.g., *-s* and *-ed*) for each content word in the window, and null tokens for words without inflections in the window, are counted for each 10-word window. This count is then divided by the total number of windows in the text to calculate within-subset variety scores for the entire text. This improves upon simply counting the number of inflectional types in the text because a simple type count would likely correlate with text length. A similar approach is used to assess derivational morpheme variety. However, since a content word could have multiple derivational morphemes, the windows of 10 words and/or null counts could have multiple derivational morphemes per word. Thus, a window of ten derivational morphemes and/or null counts may reflect 10 content words or fewer.

Morphological complexity. TAMMI 2.0 calculates an index for inflectional morphemes based on the MCI reported in Brezina and Pallotti (2019) by using the morphological variety counts above. For inflections, a between-subset diversity score is calculated. The between-subset diversity score is the average number of unique morphemes when comparing subsets where subset are the same 10-word windows found in the within-subset variety score. As an example of unique morphemes, *I loved him* and *she loves him* each have one unique morpheme, *-ed* and *-s*. The within-subset variety score is then divided by the between-subset diversity score (i.e.,. This score is then divided by the number of subsets minus 1. The same approach is followed to

produce an MCI for the derivational morphemes, which was not reported by Brezina and Pallotti (2019).

Morpheme type-token counts. TAMMI 2.0 includes indices of type-token ratios (TTR) for both inflectional and derivational morphemes. For inflectional morphemes, we use the number of unique inflectional morphemes by 10 content word window divided by the length of the window (knowing that the last window may be less than 10 words) and average the score across the text. For derivational morphemes, we calculate a similar metric, but we use a 10-morpheme window because some content words have more than one derivational morpheme.

Frequency, family size, and hapax counts. TAMMI 2.0 depends on MorphoLex to calculate variables related to frequency/length, family size counts and frequency, and hapax counts. TAMMI 2.0 matches tokens reported in spaCy to the MorphoLex dictionary. Like basic counts, TAMMI 2.0 computes mean scores for MorphoLex frequency/length, family size counts and frequency variables within a text by taking the count for each variable and dividing it by the number of content words (i.e., all words not in the spaCy stopword list) in the text to provide a normed score. Additionally, there are normed scores by the number of prefixes, suffixes, and total affixes. The MorphoLex variables calculated in TAMMI 2.0 are discussed below.

Morpheme frequency/length counts. For roots, prefixes, suffixes, and all affixes, TAMMI 2.0 extracts frequency counts for morphemes from MorphoLex. The frequency count comes from the HAL counts found in the ELP (Balota et al., 2007). TAMMI 2.0 computes a raw frequency count and a logged frequency count. TAMMI 2.0 also calculates the average length of the roots, prefixes, suffixes, and all affixes.

Morpheme family size counts. For roots, prefixes, suffixes, and all affixes, TAMMI 2.0 derives family size counts from MorphoLex. Family sizes for morphemes are calculated by

counting the number of word types to which a morpheme can attach itself. As an example, in the pool of words *attendance, pleasance, pleasure, appearance*, the suffix -*ance* has a family size of 3, but the root *pleas* has a family size of 2 (example taken from Sánchez-Gutiérrez et al., 2017). For roots, family size is calculated by the number of words a root can produce (e.g., the count for the number of words that have *theo* as a root).

Morpheme family size frequency. TAMMI 2.0 also reports the percentage of other words in the family that are more frequent (PFMF) from MorphoLex. This feature counts the percentage of morphemes per word that are more frequent by dividing the number of more frequent words in a family by the total number of family members. For instance, *word, wordiness,* and *wordlessly* all have the same root (i.e., *word*), but the word *word* is the most frequent type in the family (PFMF = 0%) whereas *wordlessly* has 10 terms that are more frequent (PFMF = 45%) and *wordiness* has 15 types that are more frequent (PFMF = 70%; example taken from Sánchez-Gutiérrez et al., 2017). Thus, a lower value indicates a word that is more frequent in the family, and higher PFMF values can contribute to greater morphological complexity.

Hapax counts. Hapaxes are defined in MorphoLex as words or roots that only appear once in a corpus. Affixes that attach to a greater number of hapaxes are more productive and can be used to create new words. TAMMI 2.0 derives two types of hapax counts from MorphoLex: the number of prefixes/suffixes/affixes that are attached to hapaxes and the number of hapaxes that include prefixes/suffixes/affixes.