Data dictionary of the MorphoLEX English database

Item-level variables

The following variables apply to the whole lexical item for each entry of the database.

ELP_ItemID

Unique ID used to identify the lexical item in the complete English Lexicon Project database. [1]

Word

The lexical item.

POS

Part-of-speech of the lexical item.

Nmorph

Number of morphemes contained in the morphological segmentation of the lexical item.

PRS_signature

Prefix-Root-Suffix signature of the lexical item. This signature is composed of 3 numbers, each indicating the number of prefixes, roots, and suffixes identified in the segmentation of the lexical item. For example, a word with two prefixes, one root and one suffix will have the following PRS signature: 2-1-1.

MorphoLexSegm

Morphological segmentation of the lexical item, where each morpheme is represented by its canonical form.

[1] Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. Behavior Research Methods, 39, 445-459.

Morpheme-level variables

The following variables were computed for each morpheme of the database. A given lexical item can contain many morphemes, and thus will have an instance of each of the morpheme-level variables for each of its morphemes.

These variables were named according to the following naming convention:

`PREF/ROOT/SUFF[position]_[variablename]`

For example, the `FamSize` variable for the second prefix of a lexical item is named `PREF2_FamSize`. In the following descriptions, any of `PREF#`, `ROOT#`, `SUFF#` (where `#` stands for a number) can replace the `X` wildcard character.

With the exception of `X_length`, *all morpheme-level variables were computed using the data available in the English Lexicon Project database.*

X_PFMF

Percentage of words more frequent than the lexical item in the morphological family of the morpheme. For example, if `PREF1_PFMF` is equal to 0, this means the lexical item is the most frequent word containing the first prefix.

X_FamSize

Morphological family size of the morpheme, i.e. the number of lexical items containing the morpheme.

X_Freq_HAL

Cumulative HAL frequency of the morpheme over all lexical items in the English Lexicon Project database. For example, if a morpheme occurs in two lexical items, with respective HAL frequencies of 10 and 3, then the `Freq_HAL` of the morpheme is 13.

X_P

First morpheme productivity score. Takes a value between 0 and 1. Intuitively, this score captures the probability, according to a given corpus, that the morpheme appears in a hapax legomenon. It is defined as:

$$P_{C_m} = \frac{H_{C,m}}{STF_m}$$

where $H_{C,m}$ is the number of hapax legomena in corpus C containing the morpheme m, and STF_m is the summed token frequency of morpheme m, i.e. its `X_Freq_HAL` (see above).

X_P*

Second morpheme productivity score. Takes a value between 0 and 1. Intuitively, this score captures the probability, according to a given corpus, that any hapax legomenon will contain the morpheme. It is defined as:

$$P^*_{C_m} = \frac{H_{C,m}}{H_C}$$

where $H_{C,m}$ is the same as in `X_P` and H_C is the total number of hapax legomena in corpus C.

X_length

Number of characters contained in the canonical form of the morpheme.