

# 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.

### **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.

## 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:

PREFIX/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.

### **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.