EwD Corpora Information

Summary

> Pre-Reading Note

Pre-Reading Note

Access corpora_info_in_ewd.md to read this article in EwD L0.

1. The Corpus Row

1.1. Row Format

EwoD | L0 ⊠ L1 < L2 > L3 ¦ Full

1.2. Explaining the Columns

EwoD stands for English without Diacritics. This our term for words or texts in pure standard English, without diacritics or any other sort of pronunciation guidance. Sometimes we may also use the term "standard English" for the same purpose.

L0, **L1**, **L2**, **L3** and **F** are the five main *variants* of English with Diacritics (EwD), the first four also representing the didactic levels. Below is the explanation for each of them:

- [TODO: complete this explanation] The L0 (Level 0) is EwD's **absolute-beginner** level. [Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.] *When omitted on a corpus row, the L1 word is equal to the EwoD word.*

- The L1 (Level 1) is EwD's **beginner** level. It includes only the indications that are essential for newcomers. It avoids the specifics of unstressed syllables and some nuances[1] of stressed syllables. *When omitted on a corpus row, the L1 word is equal to the EwoD word.*

- The L2 (Level 2) is EwD's **intermediate** level. It provides a more detailed guidance, utilizing a decent amount of applicable indications to guide learners towards a holistic understanding of English pronunciation. In this level, learners will acknowledge the patterns of unstressed syllables and explore nuances[1] of stressed syllables not covered in the previous levels. *When omitted on a corpus row, the L2 word is equal to the L1 word.*

- And the L3 (Level 3) is the **advanced** level of EwD. It is also known as the *minimal* variant. After exposure to extensive indications of the previous level, this level retains only the most essential ones. Thus, learners will understand which of EwD's indications are actually unnecessary (or not very necessary) as they gradually grasp the patterns of spoken English. *When omitted on a corpus row, the L3 word is equal to the L1 word.*

- And finally, the Full variant is the **most phonetic** presentation of EwD. It is not a level, but a variant meant to be used on *dictionary entries* or the like. As such, it goes very specific and ships all applicable marks, with the exception of a few beginner level (L0 or L1) ones, as well as some marks deemed very unnecessary in any EwD variant. Besides dictionary entries, this variant is suitable for lyrics and other types of content when highlighting the specific pronunciation patterns being used is desired. *When omitted on a corpus row, the Full variant word is equal to the L2 word.*

2. The Accept-or-Except Strategy (AES)

The Accept-or-Except Strategy (AES) in EwD is a key mechanism that balances the integrity of standard English spelling with the need for phonetic clarity. This strategy primarily focuses on *retaining standard spellings*[2], and on applying *diacritical marks* only when necessary to assist in pronunciation — this is the **Accepted side**. Conversely, when standard spellings obscure the correct pronunciation, and diacritics would complicate them more than clarify, EwD shifts to the **Exception side**, adopting so-called *fallback graphemes* that more accurately reflect spoken English.

This approach eases the learning process and facilitates the transition from EwD-assisted reading to independent reading. By preserving standard spellings as much as possible, EwD helps learners recognize and understand the spelling-to-speech patterns of English, thereby gradually building their confidence and proficiency in pronouncing English words without diacritical assistance.

2.1. AES Sides

Below are the two sides of the just-explained strategy:

- Accepted side — The collective of untouched words and words that are treated with diacritic(s). About **97%** of the 1000 most common English words are on this side. In the context of the AES, we may also refer to this as the "A" side.

- Exception side — The collective of words for which diacritics are deemed not enough or not suitable to indicate pronunciation. The words from this side are known as *exceptions*. About **3%** of the 1000 most common English words are on this side. In the context of the AES, we may also refer to this as the "E" side.

2.2. The "Not Disclosed / Disclosed" Approach

[TODO: continue this explanation] For exceptions, the EwD material employs a "not disclosed / disclosed" approach to hide or clarify their pronunciation, according to the necessity. The "not disclosed" form is a given exception with the irregular grapheme preserved, while the "disclosed" form is the same word with such grapheme replaced by its corresponding *fallback grapheme*. [Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.]

3. Legend & Row Tags

[TODO: add explanation]

3.1. Pronunciation Variants

[TODO: add explanation]

- str: stressed
- un: unstressed
- GA: General American accent
- GB: General British accent (*aka* Received Pronunciation – RP)
- CP: Common Pronunciation
- LCP: Less Common Pronunciation
- SP: Spelling Pronunciation
- AP: Academic Pronunciation

3.2. Word Groups

[TODO: add explanation]

-

-

- [A]: Accepted — Words on the **Accepted side**; the collective of words from [Au] and [Ad] subgroups. The words from this group are known as *accepted word.*

    - [Au]: Accepted, untouched — [A] words that *don't* receive diacritics in any level; the so-called *untouched words.*

    - [Awd]: Accepted, with diacritics — [A] words that receive *diacritic(s) in at least one level*[2].

    - [Awd^n]: Accepted, with diacritics, n — [Ad] words that receive *diacritic(s) only at given level(s).* "n" corresponds to the level number(s).

- [E]: Exception — Words on the **Exception side**; the collective of words from [Ew] and [Ep] subgroups. The words from this group are known as *exceptions.*

    - [Eap]: Exception, all pronunciations — [E] words whose *all* used pronunciations require fallback graphemes to be clarified. In the corpora these words are always tagged accordingly.

    - [Eap^n]: Exception, all pronunciations, n — [E] that are treated with fallback graphemes only at given levels. "n" corresponds to the level number(s). In the corpora these words are always tagged accordingly.

    - [Esp]: Exception, some pronunciations — [E] words whose *some* of its used pronunciations (but now all of them) require fallback graphemes to be clarified. In the corpora these words are always tagged accordingly.

3.3. Tags about standard English spelling

[TODO: add explanation]

- AS: American spelling
- BS: British spelling
- FE: Word with *false etymology*

Appendix

Notes:

1. nuance ... [1](GB/GA) [2](CP/LCP)

2. In fact, EwD strives to keep as meny{many} words as possible on this side, with about 97% of the 1000 (one thousand) most frequent English words falling under the **Accepted** category in EwD v2.

Abbreviations:

- aka: also known as

Glossary: (TODO: add definitions)

- corpus:

- corpora: