Processing stress, length marks and some diacritics

Wilbert Heeringa

April 16th, 2024

Stress

Before applying the Levenshtein distance to the transcriptions, it is checked whether a stress mark is always found right before a vowel. In cases where that is not the case, the stress mark is moved to before the next vowel.

When running the Levenshtein distance, primary stress will normally be processed as an indel (insertion or deletion) and weighs as maxIndel (i.e. the highest weight that an indel can have). Secondary stress will normally only occur as indel too, but it weighs as maxIndel * 0.5. A match between a stress mark and a speech segment is theoretically possible but unlikely.

A substitution of primary stress for secondary stress and vice versa is also possible. A substitution weighs as 0.5 * maxIndel.

Length

Before calculating the Levenshtein distances length marks are processed by adapting the phonetic transcriptions as follows:

- if a segment is transcribed as extra short (e.g. [ă]), it remains unchanged;
- if a segment does not have any length mark, it is doubled, e.g., [a] becomes [aa];
- if a segment is marked as half-long, it is tripled, e.g., [a·] becomes [aaa];
- if a segment is marked as long, it is quadrupled, e.g., [aː] becomes [aaaa].

As to diphthongs, e.g. [au] becomes [aauu], [a·u] becomes [aaauu], etc.

As to affricates: IPA [ʧ] is processed as half a [t] and half a [ʃ], IPA [ʤ] is processed as half a [d] and half a [ʒ]. Here half sounds are the same as extra short sounds.

Some diacritics

The diacritics that are processed are listed in table below. The processing is as follows. Assume an aspirated sound such as [tʰ] is compared to [s]. Then the [tʰ] is treated as half a [t] and half an [h]. The distance between [t] and [s] is calculated and the distance between [h] and [s] is calculated, and subsequently the average of the two distances is calculated as the final distance. For the other diacritics the same approach is applied. Here half sounds are the same as extra short sounds.

Diacritic	IPA mark	Example	Processed as short
aspirated	ʰ	tʰ	h
labialized	ʷ	tʷ	w
palatalized	ʲ	tʲ	j
velarized	ˠ	tˠ	ɣ
pharyngealized	ˤ	tˤ	ʕ
nasalized	̃	ã	n

References

Heeringa, Wilbert & Gooskens, Charlotte & Van Heuven, Vincent (2023). Comparing Germanic, Romance and Slavic: Relationships among linguistic distances. *Lingua* 287, 103512.

Heeringa, Wilbert & Van Heuven, Vincent & Van de Velde, Hans (2022). *LED-A: Levenshtein Edit Distance App* [Computer program]. Retrieved 2 January 2023 from https://www.led-a.org.