

Codebook for the OSF file “Investigating the Syntax-Discourse Interface in the Phonetic Implementation of Discourse Markers” (Hutin et al. 2023)

The dataset is based on the LOCAS-F corpus (Degand et al. 2014). The acoustic measurements were made using a Praat script adapted from Lancien (2021).

Each line corresponds to one /a/ or /k/ token uttered in a discourse marker.

The columns correspond to the following information:

name_textgrid_file: Name of the Textgrid file used for the acoustic analysis. The name of each textgrid was different for each speaker of the corpus, so each name corresponds to a unique identifier for each speaker in each audio file.

label: Phoneme /a/ or /k/.

start_time: Time of the beginning of the phone utterance.

end_time: Time of the end of the phone utterance.

duration: Duration of the phone utterance.

previousLabel: Previous phone in the utterance.

followingLabel: Following phone in the utterance.

mean_F0(Hz): Mean value of F0 of the phone (in Hertz)

min_F0(Hz): Minimum value of F0 of the phone (in Hertz)

max_F0(Hz): Maximum value of F0 of the phone (in Hertz)

std_dev_F0(Hz): Standard deviation of the mean values for F0 (in Hertz)

mean_F1(Hz): Mean value of F1 of the phone (in Hertz)

min_F1(Hz): Minimum value of F1 of the phone (in Hertz)

max_F1(Hz): Maximum value of F1 of the phone (in Hertz)

std_dev_F1(Hz): Standard deviation of the mean values for F1 (in Hertz)

mean_F2(Hz): Mean value of F2 of the phone (in Hertz)

min_F2(Hz): Minimum value of F2 of the phone (in Hertz)

max_F2(Hz): Maximum value of F2 of the phone (in Hertz)

std_dev_F2(Hz): Standard deviation of the mean values for F2 (in Hertz)

mean_HNR(dB): Mean value of Harmonic-to-Noise Ratio of the phone (in decibels)

min_HNR(dB): Minimum value of Harmonic-to-Noise Ratio of the phone (in decibels)

max_HNR(dB): Maximum value of Harmonic-to-Noise Ratio of the phone (in decibels)

std_dev_HNR(dB): Standard deviation of Harmonic-to-Noise Ratio of the phone (in decibels)

syll: Transcription in International Phonetic Alphabet (IPA) of the syllable in which the phoneme appears.

previous_syll: Transcription in International Phonetic Alphabet (IPA) of the syllable preceding the one in which the phoneme appears.

next_syll: Transcription in International Phonetic Alphabet (IPA) of the syllable following the one in which the phoneme appears.

word: Transcription in Latin orthography of the word in which the phoneme appears.

previous_word: Transcription in Latin orthography of the word preceding the one in which the phoneme appears.

next_word: Transcription in Latin orthography of the word following the one in which the phoneme appears.

pos_dismo: Part of speech of the word in which the phoneme appears, as identified by DisMo (Christodoulides et al. 2018).

previous_pos_dismo: Part of speech of the word preceding the one in which the phoneme appears, as identified by DisMo (Christodoulides et al. 2018).

next_pos-dismo: Part of speech of the word following the one in which the phoneme appears, as identified by DisMo (Christodoulides et al. 2018).

speaker: ID of the speaker within the original audio file.

genre: Type of discourse

formality: Whether the discourse is formal, or informal.

nb_of_speakers: Number of speakers in the interaction.

situation_type: Whether the interaction is a monologue, dialogue or multilogue.

interaction: Whether the situation is interactive, semi-interactive or non-interactive.

preparation: Whether the speech is prepared, semi-prepared or unprepared.

media: Whether the speech is broadcasted in the media (e.g., political debates) or not (e.g., scientific conference), or hybrid (e.g., homilies, political speeches...).

professionalism: Whether the interaction takes place in a professional or a private setting.

naturality: Whether the speech is spontaneous (natural) or semi-elicited.

year_of_recording: Year of the recording.

speaker_gender: Gender of the speaker: F for female, M for male.

speaker_nationality: Nationality of the speaker: BE for Belgian, FR for French, CH for Swiss.

pos_dismo_large: Whether the word in which the phoneme appears is an adverb (Adv), a coordinating conjunction (CC), a subordinating conjunction (SC) or an interjection (Ij).

References.

- L. Degand, L. J. Martin, and A.-C. Simon, "Unités discursives de base et leur périphérie gauche dans LOCAS-F, un corpus oral multigenres annoté" SHS Web of Conferences, vol. 8, 2613-2625, 2014. [Online]. Available: <https://doi.org/10.1051/shsconf/20140801211>
- G. Christodoulides, M. Avanzi, and J. Goldman, "DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator. An Evaluation on a Corpus of French Spontaneous and Read Speech" CoRR, vol. abs/1802.02926, 2018. [Online]. Available: <http://arxiv.org/abs/1802.02926>
- Hutin, M., L. Degand, and M. Allasonnière-Tang, "Investigating the Syntax-Discourse Interface in the Phonetic Implementation of Discourse Markers" Proceedings of Interspeech 2023, In Press.
- M. Lancien, "Le rôle de la réduction phonétique dans l'expression de la proximité sociale : Étude acoustique des voyelles orales du français québécois dans différentes situations de communication." Thèse, Université de Lausanne (Suisse), May 2021. [Online]. Available: <https://hal.science/tel-03709234>