Corpus data metadata in the EXMARaLDA Demo corpus

Communication

Description

- **Communication type:** Open vocabulary of communication types including medium.
- **Project name:** The name of the project responsible for the corpus.
- **Background information (optional):** Any text describing the communication situation or its background further.
- **Source (optional):** The source of published recordings.

Location [Type = Communication]

The main location where the communication took place. (See Location)

Language [Type = Communication]

All languages used in the communication. (See Language)

Recording

Recording *<Name>* (*<Recording duration>*)

File: Relative path to an individual media file.

Transcription

The transcription is described further within the segmented transcription description containing the elements listed below. The description for the basic transcription only contains the "segmented" and "File" elements.

Description

- **Transcription name:** The name of the transcription.
- **Transcription convention:** The transcription conventions used.
- **Alignment status:** Indicates whether the transcription is fully, partly or not aligned with the (wav) recording.
- **Transcription status:** Indicates whether the (wav) recording has been fully, partly or not transcribed.
- **Transcription date:** The date the transcription was created.
- **Transcriber:** The person who created the transcription.
- **Segmentation algorithm:** The automatic segmentation algorithm used by the EXMARaLDA system to create the segmented transcription from the basic transcription.
- **File:** The relative path to the transcription file.
- **segmented:** {true, false} Indicates whether this is a segmented (true) or basic (false) transcription.
- **EXB-SOURCE:** The basic transcription from which the segmented transcription was created.
- **sc:** The number of segment chains in the segmented transcription.
- **e:** The number of events in the segmented transcription.
- **HIAT:u:** The number of HIAT utterances in the segmented transcription.
- **HIAT:w:** The number of HIAT words in the segmented transcription.
- **HIAT:non-pho:** The number of HIAT non-phonological elements in the segmented transcription.
- **HIAT:ip:** The number of HIAT punctuation elements in the segmented transcription.

Location

Locations contain information about time and place. Various types of information on a speaker that can be associated with a certain time and place are therefore encoded as locations. Depending on the information type, different elements are required or optional. The type element is always required.

Type: This attribute is used to differentiate between various location types, such as communication locations or locations related to education, occupation, residence or stays of a certain speaker.

Street: The street name (and house number).

City: The city.

PostalCode: The postal code.

Country: The country, using the ISO Codes for the Representation of Names of Languages.

Period start: The start date and time encoded as DD.MM.YYYY HH:MM.

Exact?: Indicates whether the date and time indicated is exact. Non-exact information can be used where exact information is not available or where exact information would prevent anonymization of participants but approximate information is still necessary. To specify how approximate a non-exact information is, or to encode the required birth date of a participant as unknown, the Key **Precision** is used in the location's description.

Duration: The duration as <years> <days> <hours> <minutes> <seconds> <milliseconds>. Months can't be encoded as such, since they vary in number of days and therefore couldn't be compared.

Key/Value: The description can be used for further information. The key **Region** has been used to describe any area bigger or smaller than a city.

Language

ISO 639-3 Code: The ISO code for the language. (Or XXX if the language has no code.)

Name: The name of the language will be generated from the ISO code, but if the ISO code is not known, typing the language name in English in this field will make a pick list of possible language names appear.

Type: This optional attribute is used to differentiate between various language types, such as a language used in a communication or on the other hand L1s or L2s of a certain speaker. Due to the complex linguistic reality when it comes to multilingualism, first and second languages and their definition, we only differentiate between L1s and L2s where obvious, and leave the type attribute of languages without value where the situation is not clear.

Key/Value: The description can be used for further information on language variety, proficiency, speech characteristics etc.

Speaker

Description

The description allows any Keys and values. Those listed below are used systematically in the Demo corpus.

- **Name:** The participants name.
- **Function:** "subject" (for HZSK Core Metadata Compliance).
- **Family: Children (optional):** Information on the participant's children.

- **Family: Marital status (optional):** Information on the participant's marital status.
- **Family: Profession father (optional):** Information on the participant's father's profession.
- **Family: Profession mother (optional):** Information on the participant's mother's profession.
- **Family: Siblings (optional):** Information on the participant's siblings.

Much information on speakers is encoded as locations. For each piece of information, such as different educations, one location element is used. Below the required elements of the location itself and systematically used elements of the location's description are listed.

Location [Type = Birth]

This location type stores information on the participant's birth.

Country: The country, using the ISO 3166-1 encoding list of the countries.

Period start: The start date and time encoded as DD.MM.YYYY HH:MM. If the birth date is unknown, this is set to 01.01.1900 00:00 and the **Precision** key is used to describe this (Value = birth date unknown).

Exact?: Indicates whether the date and time indicated is exact.

Location [Type = Education] (optional)

This location type stores information on the participant's education.

- **Name (optional):** The type of school or the name of a specific school.
- **Degree (optional):** The degree obtained.
- **Subject (optional):** The subject(s) of the education.

Location [Type = Occupation] (optional)

This location type stores information on the participant's occupation.

- **Name (optional):** The name of the occupation.

Location [Type = Residence] (optional)

This location type stores information on the participant's residence.

Location [Type = Stay] (optional)

This location type stores information on places where the participant has been staying for shorter periods, for example when traveling.

Language

All speakers are required to have at least one language.