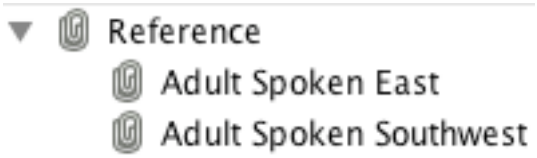
Understanding Coma Metadata

Structure of Coma Metadata

Coma-Metadata consists of 5 predefined data-„containers“: corpora, communications, speakers, transcriptions and recordings. Besides, there are further datatypes that can exist inside these containers. It is important to understand the connection between these containers and datatypes.

Corpora:

Corpora are the top-level containers for all other containers and datatypes. They can contain either Speakers and Communications or further Corpora, but not both.



The example shown consists of one corpus („Reference“) that holds two subcorpora. The „Reference“-Corpus cannot contain actual corpus data (= speakers and communications), since it already contains other corpora.

Communications

Communications are used to manage the events where the transcribed conversations took place. Communications typically feature speakers and there can be recordings and transcriptions of the conversation. In the coma data-model, recordings, transcriptions and speakers are linked to communications. Furthermore, all things noteworthy of the communicative situation (time, place and circumstances, languages spoken) are stored with the communication.

Speakers

Speakers are – as the name suggests – the persons that participate in the communication. Speakers don't have to be actual persons (automated dialog-systems also qualify) and they don't have to actually speak – as long as they are important for understanding what is happening in the conversation, they should be registered.

The speaker-datatype should contain everything that is important about that speaker, like date and time of birth, language learning history etc.

Since speakers can be linked to multiple communications, data that is only relevant for one communication should not be saved with the speaker, but with the communication.

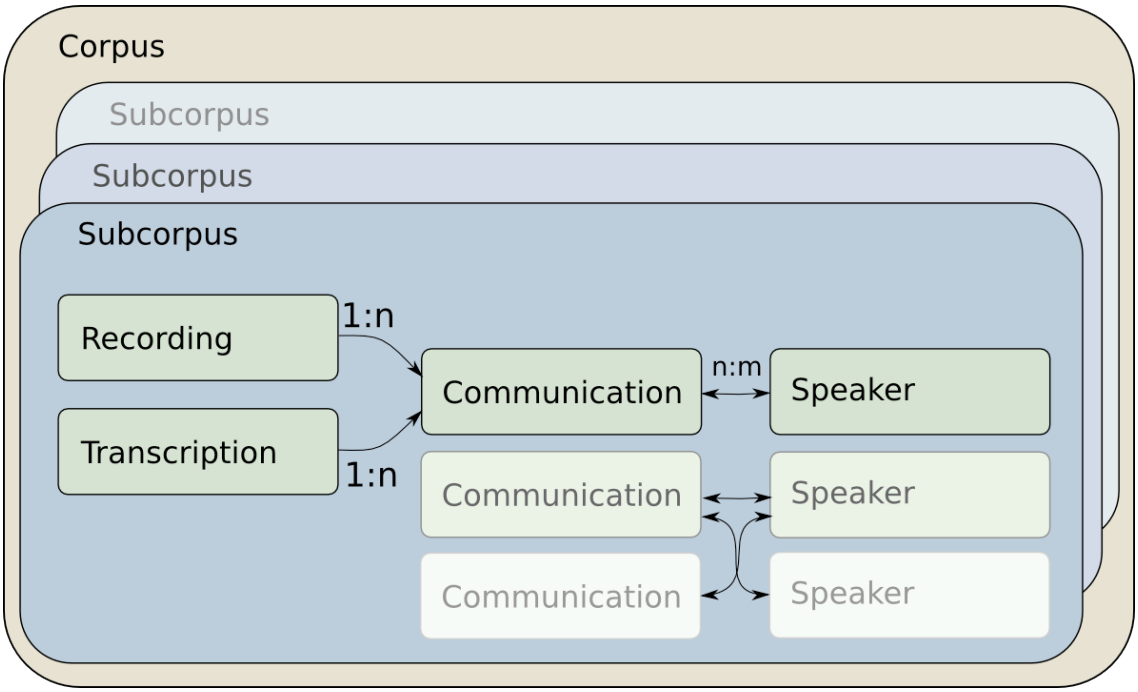
Recordings

Recordings reference to the actual (audio or video) recording of the communication. Recordings are always connected to a communication and can not exist on their own.

Transcriptions

Transcriptions establish the link to actual EXMARaLDA-transcription-files. Coma manages basic- as well as segmented transcriptions. There is an option inside the Coma-preferences-panel to toggle whether basic transcriptions are to be shown or not, since tools like the EXMARaLDA search tool "EXAKT" only handle segmented transcriptions.

Like recordings, transcriptions must always be linked to a communication and cannot be managed without them. They are also linked to their recordings through communications.

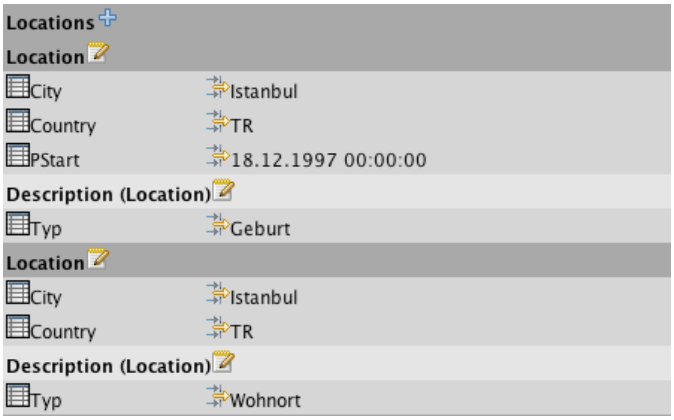


Further Datatypes

To capture actual metadata, further datatypes exist. Two of them are of special importance:

Location

Locations represent a location *at a certain time*.



A location does not have to hold place and time information, but it can: In that given example, one location encodes birth date and location of a speaker, the second location encodes only the location of residence. It is important to remember to use locations even if one only wants to register the time of a special event.

Description

Since it is extremely difficult to find a standardized set of metadata-fields for all areas of research, most of the metadata in Coma is encoded through free key-value pairs. These pairs are collected inside Description-fields. These exist in all Coma-datatypes: There can be descriptions for corpora, for communications, for recordings etc.

The example shows a descriptions for one speaker. Since the keys inside descriptions can be named freely, it is very important to create a unified vocabulary of description-keys for corpus metadata. Coma's templates can help to harmonize descriptions and simplify their input.