TALP Tourism Dialogues - Catalan

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

The "TALP Tourism Dialogues - Catalan" comprises the recordings of 160 dialogues (8,600 turns) in the touristic domain. Each participant took the role of either customer or agent and talked spontaneously to achieve a predefined goal defined in a given scenario. The scenarios include hotel, travel agency, tourism information office and railway/airline company. The data was recorded over the telephone (a-law, 8kHz) using a platform that imposed half-duplex communication: there are not turn overlapping. The total recordings time is 22 hours. The database includes the orthographic transcription enriched with additional labels to indicate external noises, speaker noises and disfluences in spontaneous speech.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Nuria Castells
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 4137856
Fax: +34 93 401 6447
e-mail: castell@lsi.upc.edu

2.2 . Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3 . Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to Universitat Politècnica de Catalunya. The resource is free, license-based, for research purposes and fee license-based for commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The archive that will be uploaded on the MetaShare platform will contain several files containing documentation (copyright, readme, documents, …), speech files (one file for each turn) and label files (each speech file has an accompanying label file) in a nested structured directory..

3.2. Encoding

Documentation is encoded in plain text.

Speech files are stored in a riff (wave) file with a header and as a sequences of 8-bit 8 kHz A-law uncompressed speech samples (CCITT G.711 recommendation). Each dialogue is in a separate directory and each turn is store in a separate file.

Each speech file has an accompanying SAM label file.
Label files contain information about the database, speech signal coding, speaker code, segmentation and enhanced orthographic transcription

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains about 8.600 files (turns) of recorded speech (22 hours)

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

The database was created for collecting spontaneous dialogue data in the tourism domain. The data can be used to study dialog and spontaneous speech. It can also be used for train domain-specific acoustic and language models for ASR systems in Catalan.

4.2. Content description.

The contents of the database are spontaneous dialogs in the tourism domain. Two volunteers talk to each other using a telephone platform. Therefore, all the iteration is by voice. One speaker acts as the agent while the other plays the role of client. They are provided with an scenario (goal) and some tools related with the task

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The transcription included in this database is an orthographic, lexical transcription with a few details that represent audible acoustic events (speech and non speech) present in the corresponding waveform files. The annotation also marks some disfluences as re-start, end of word missing, etc. The extra marks contained in the transcription aid in interpreting the text form of the utterance. Transcriptions were made in two passes: one pass in which words are transcribed, and a second pass in which the additional details are added.

4.4. Lexicon. Description of the lexicon (if applicable)

The lexicon file is an alphabetically ordered list of distinct lexical items (i.e. splitting tokens at every space, essentially words in our case) which occur in the corpus with the corresponding pronunciation information. Each distinct word has a separate entry. As the lexicon is derived from the corpus it uses the same alphabetic encoding for special and accented characters as used in the transcriptions. The pronunciation information has been produced semi- automatically using the in-house UPC grapheme-to-phoneme system. Special care was put in proper names to identifiy the language of the names and transcribe according to that language. The number of entries is around 106k

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

There were no specific speaker selection strategy but the accent and gender of the speakers was registered. The number of speakers is 58.

4.6. Recording platform

The database was recorded using a telephone platform called GAIA, developed at UPC. The *client* called the platform to ask for an *agent.* The *agent* is connected with the *client* and the speech and other information is logged in the platform.

The platform is half-duplex: each speaker needs to push '#' to give the turn to the other speaker.

Most of the calls with telephones at UPC. Therefore, they represent the internal network.

The platform used digital telephone lines: the speech is recorded directly in the original A-LAW enconding..