3/24 BN (Catalan BN)

1. BASIC INFORMATION

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

The 3/24 BN database contains 80 hours of recordings of the Catalan news television channel 3/24. 19 hours of this database are fully transcribed while the remaining data are solely segmented and annotated with respect to speaking style, recording condition and speaker. The transcription follows the general guideline generated within the TC-STAR project for European Parliament Plenary Sessions but it was extended to include additional information as the language, background condition, silence/voice segmentation, speaker segmentation and acoustic events featuring additional time stampsThe transcriptions have four layers. Transcriptions follow the TRS format produced by the Transcriber transcribing tool. .

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: José A. R. Fonollosa
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 4016439
Fax: +34 93 401 6447
e-mail: jose.fonollosa@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

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, …), 20 speech files (4 complete hours per file), each having an accompanying video, and transcription file) in the same directory.

3.2. Encoding

Documentation is encoded in word/pdf/plain text.

The recorded video files are stored as MPEG-2. Whereas the extracted speech files are stored as 16-bit 48 kHz uncompressed speech samples .

Each speech file has an accompanying transcription file.
Transcription files contain information about the database, speech signal coding, speakers (public figures identified by names, others by their role within the broadcast, e.g. translator, guest), turns, segmentation, speaking style, channel, background sounds, acoustic events and literal transcriptions of the speech.

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

The corpus contains about 20 speech files (audio and video)/80 hours of recorded speech and needs about 34 GB for disk storage.

4. CONTENT INFORMATION

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

This is a database for training ASR systems in Catalan, but also contains minor proportions of Spanish.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The contents of the database are: Complete broadcast news sessions including interviews, reports from different recording environments, segmented in speaker turns, phonetically rich, read and spontaneous speaking style.

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 extra marks contained in the transcription aid in interpreting the text form of the utterance. Transcriptions were made in three passes: one pass in which speaker segments and environmental conditions are added, a second pass adding acoustic events and their time stamps and, a third pass transcribing those segments not featuring music and speech overlap .
Transcriptions were checked periodically for quality control. Two persons transcribed samples of the same transcription task and results were compared.

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

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

The database recordings contain segments from 1599 adult Catalan speakers of unknown accent (981 male, 421 female, 197 unknown), and 351 adult Spanish speakers (217 male, 82 female, 52 unknown) Speakers are unbalanced in gender favouring male speakers in total duration.

4.6. Recording platform

4.6.1.Domain(s), environments,

Four recording environments were defined:

Studio: segments originate from speakers located in the studio.

Telephone: segments originate from speakers over a telephone.

Outside: segments originate from speakers outside of buildings, e.g. on streets, public space

None: segments that have non of the above classification.

4.6.2 Recording platform

The recordings originate from DVB-T video streams, whereas the audio channel is provided with 48 kHz sample rate, 16 bit uncompressed samples. The video streams are MPEG-2 encoded.