TALP Tourism Dialogues - Translation

1. BASIC INFORMATION

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

This corpus consist of the translation into English and either Spanish or Catalan of the transcriptions in the ”TALP Tourism Dialogues – Spanish” and in the ”TALP Tourism Dialogues – Catalan” databases. The dialogues were recorded over the telephone network. Each participant took the role of either customer or agent and were provided with a scenario describing the goal of the conversation. The spoken dialogs were transcribed and the translated into two other languages so that a tri-lingual corpus was produced.

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

3. TECHNICAL INFORMATION

3.1. Directories and files

The archive uploaded on the MetaShare platform contain two files containing documentation (copyright, readme), and for each language (English, Catalan, Spanish) there are two files, one with the

transcriptions or translations of the Spanish dialogs and the other with the transcriptions or translations of the Catalan dialogs. Each file contains one line per turn, with a code to identify dialog, turn in the dialog, scenario code, speaker code, language and source language.

3.2. Encoding

Documentation is encoded in text.

Translation files are encoded using ISO-8859.

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

The corpus contains about 19,000 turns translated into three languages. The size of the uncompressed database is less than 10MB.

4. CONTENT INFORMATION

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

The primary use of the database is for supporting the research on machine translation of spoken language and for training statistical machine translation models.

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 spontaneous dialogs in the tourism domain. Two volunteers talk to each other using a telephone plataform. 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.

Before translation, the transcriptions were cleaned up. However, the spontaneous style was preserved. The translation tried to be literal but correct in the target language. It was done by professional translators

4.3. 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 total number of speakers recording either Spanish or Catalan dialogs was 122.