

The Nijmegen Corpus of Casual Spanish

Francisco Torreira & Mirjam Ernestus

October 30, 2009

1 Introduction

Spanish is one of the best documented languages in the world. However, no large corpus of casual Spanish suitable for detailed phonetic analysis is available to our knowledge. The goal of this article is to introduce the Nijmegen Corpus of Casual Spanish (NCCSp from now on), a new corpus designed to fill this gap. The corpus was designed taking the Nijmegen Corpus Casual French as a model [Torreira et al., in press], which was also collected in our lab. The uniqueness of the NCCSp can be characterized as follows:

- It contains around 30 hours of casual conversations among groups of friends. This makes it possible to study a wide range of phenomena characteristic of casual speech.
- It contains speech from 52 native Madrid Spanish speakers sharing a similar educational background.
- It contains large amounts of data for every speaker (around 90 minutes of recorded speech for every group of three speakers). This allows researchers to study within-speaker variability.
- It is orthographically annotated.
- It contains video as well as audio data, which can be used by researchers interested in the use of facial and body gestures during verbal communication.

The following sections provide a detailed description of the creation and transcription of the NCCSp.

2 Corpus creation

2.1 Participants

The corpus creation was begun in March 2008. A group of university students were hired at the Universidad Politécnica de Madrid as confederates. These confederates were interviewed and asked to find two friends willing to participate in recordings of natural

conversations. These friends are referred to as *speakers* from now on. Every recording consisted of a conversation among three participants: a confederate and two speakers. All participants complied with the following conditions:

- They knew the two other participants in the recording well.
- They were of the same sex as the two other participants in the recording.
- They were university students in Madrid.
- They had been raised in the Madrid region.
- They reported not suffering from any pathology related to speech or hearing.

Speakers were invited to act as a confederate in a later recording. For this reason, nine participants took part in more than one recording session (first as a speaker and later as a confederate). In total there were 52 participants (27 female and 25 male). All participants were university students aged between 19 and 25. More details about the participants' background can be found in the NCCSp corpus package.

2.2 Recording set-up

The recording booth was sound-attenuated and had an approximate size of 4 x 2 m. The participants sat on chairs around a table. The confederate always sat on the south side of the table, while the speakers occupied the chairs on the north and west sides. The speakers were recorded on a Edirol R-09 solid-state stereo recorder. Each speaker was recorded in a separate channel. The confederate was directly recorded on a computer via a dedicated sound card. All participants wore a Samson QV head-mounted unidirectional microphone. The microphones were placed at an average distance of 5 cm from the left corner of the speakers' lips. The sampling rate used was 44.1 KHz, and quantization was set to 32 bits.

The conversations were filmed using a Sony HDR-SR7 video camera. The camera was placed in a corner of the recording room in a position that allowed us to film the two speakers, but not the confederate. In order to avoid inhibiting the speakers, we tried to make them believe that the camera was turned off during the recordings. As a first step, a small piece of duck tape was placed on each of its lights. Additionally, an unplugged cable was left hanging from the camera in order to reinforce the impression that it was turned off. Finally, we placed several unused objects near the camera, including old boxes and cables, a computer screen, several loudspeakers and other audio equipment.

2.3 Recording procedure

The recording procedure was similar to that employed during the collection of the Nijmegen Corpus of Casual French and the Nijmegen Corpus of Casual Czech <http://>

mirjamernestus.ruhosting.nl/Ernestus/. Previous research has shown that this procedure is successful at eliciting casual spontaneous speech [Torreira et al., in press]. This subsection describes the recording session in more detail.

Preparations: Confederates arrived at the Universidad Politécnica de Madrid for an interview with the first author (FT from now on) thirty minutes earlier than their friends. During this interview, FT informed the confederates that it was their responsibility to elicit natural speech from their friends, by raising appropriate topics whenever the conversation seemed to approach a dead end. In order to maximize the amount of recorded speech from the speakers, they were instructed not to monopolize the conversation. They were also informed that the conversation would be filmed, and where to sit so that only the other participants would appear in the film. Importantly, they were asked not to unveil any of these details to their friends until the end of the recording, and to pretend that they had never met FT. Finally, they were briefly instructed about an activity planned for the third part of the recording (see below for details).

At the end of the interview, the confederates were asked to wait for the other participants in the entry hall. At the time of the appointment, FT met the three participants there and asked them to wait while he made an urgent phone call. He then returned to the recording room, started the video recording, turned off the lights and closed the door. Back at the entry hall, he invited the participants to follow him to the recording room, making sure that the confederate would be the first person to enter in order to prevent the other participants from taking their seat. Once in the room, the participants were asked to stay seated and not to touch their microphones or play with any other object (e.g. keys, watch) during the conversation.

Part 1: After adjusting the recording volume during the first two minutes of the conversation, FT entered the recording booth and informed the participants that the confederate's microphone was not working properly. He then asked the confederate to come out of the room in order to try a new one. At this moment, the speakers left in the room did not know with certainty whether they were being recorded. It was precisely then that the recording was started. This situation elicited very natural speech right from the beginning of the recording.

Part 2: After a period of ten to thirty minutes (depending on the liveliness of the conversation), confederates were asked to go back into the room. The conversation then held by the three participants constituted the second part of the recordings. No instructions were provided about the topics to be discussed during this part of the conversation. Among the conversation topics addressed by the speakers during this part were exams, parties, and travel plans. Words characteristic of such topics are therefore well represented in this part of the recordings (e.g. *estudiar* 'to study' and morphologically related words $n = 86$; *viaje* 'travel' $n = 43$; *beber* 'to drink' and morphologically related words $n = 84$)

Part 3: After a period of thirty to forty minutes, FT entered the room and provided the participants with a sheet of paper describing the activity for the remaining part of the recording session. The participants were asked to choose at least five questions about

political and social issues from a list, and then negotiate a unique answer for every question. In order to encourage them to negotiate common stances rather than just discuss the chosen topics, we informed them that they would have to write down their answers at the end of the recording session. A characteristic of the speech elicited during this part is that its vocabulary reflects the chosen questions. For instance, the word *fumar* ‘to smoke’ is very frequent in this part of the recordings ($n = 217$) because most groups of participants chose to discuss a question about a recent smoking ban in Spain.

At the end of the recording, we revealed our procedures to the participants. We paid 30 euros to each of the speakers and 45 euros to the confederate as a compensation for their time. We then handed them a consent form agreeing to the use of the recordings for academic and scientific purposes. All of the participants signed the consent form.

3 Orthographic transcription

The corpus was orthographically transcribed in Barcelona by Verbio Speech Technologies S.L. using TRANSCRIBER software [Barras et al., 2001]. The transcription process consisted of three passes. In the first pass, the speech of every pair of speakers was orthographically transcribed in a two-tier annotation file using stereo-channel audio streams. Confederates, who had been recorded in a separate mono channel, were transcribed separately in a one-tier annotation file. The transcribed text is organized into chunks corresponding to not more than 15 seconds of the speech signal. In the second pass, non-speech events (e.g. laughter, filled pauses, etc) were added to the orthographic transcription, the location of chunk boundaries was readjusted, and the spelling of the transcription was checked using the *Diccionario de la Real Academia Española* (<http://www.rae.es/rae.html>) as a reference. In the third pass, an automatic revision of the formatting of the transcription files was performed. Every pass was carried out by a different transcriber.

The orthographic transcription of the corpus contains around 393 000 word tokens and 16 500 word types (distinct orthographic forms) distributed over 98 000 chunks. Part 1 contains around 83 000 word tokens, while Parts 2 and 3 contain each around 155 000 word tokens.

4 Corpus availability

Information about how to obtain a copy of the corpus can be found online at <http://mirjamernestus.ruhosting.nl/Ernestus/NCCSp>. This webpage also provides audio and transcription examples, scripts for searching the corpus using Praat, and more information about each participant and conversation in the corpus.

References

- [Barras et al., 2001] Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2001). Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1-2):5–22.
- [Torreira et al., in press] Torreira, F., Adda-Decker, M., and Ernestus, M. (in press). The Nijmegen Corpus of Casual French. *Speech Communication*.