

The Nijmegen Corpus of Casual French

Francisco Torreira, Martine Adda-Decker, Mirjam Ernestus

▶ **To cite this version:**

Francisco Torreira, Martine Adda-Decker, Mirjam Ernestus. The Nijmegen Corpus of Casual French. Speech Communication, 2010, 52 (3), pp.201. 10.1016/j.specom.2009.10.004 . hal-00608402

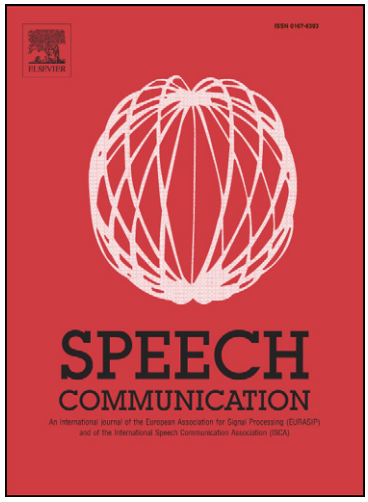
HAL Id: hal-00608402
https://hal.science/hal-00608402

Submitted on 13 Jul 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Accepted Manuscript



The Nijmegen Corpus of Casual French

Francisco Torreira, Martine Adda-Decker, Mirjam Ernestus

PII: S0167-6393(09)00162-9
DOI: 10.1016/j.specom.2009.10.004
Reference: SPECOM 1840

To appear in: *Speech Communication*

Received Date: 18 May 2009
Revised Date: 16 October 2009
Accepted Date: 16 October 2009

Please cite this article as: Torreira, F., Adda-Decker, M., Ernestus, M., The Nijmegen Corpus of Casual French, *Speech Communication* (2009), doi: 10.1016/j.specom.2009.10.004

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

The Nijmegen Corpus of Casual French

Francisco Torreira[a,b], Martine Adda-Decker[c], Mirjam Ernestus[a,b]

[a] *CLS, Radboud Universiteit, Wundtlaan 1, 6525 XD, Nijmegen, The Netherlands*
[b] *Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD, Nijmegen, The Netherlands*
[c] *Spoken Language Processing Group and Situated Perception Group, LIMSI-CNRS, BP 133 91403 Orsay Cedex, France*

Abstract

This article describes the preparation, recording and orthographic transcription of a new speech corpus, the Nijmegen Corpus of Casual French (NCCFr). The corpus contains a total of over 36 hours of recordings of 46 French speakers engaged in conversations with friends. Casual speech was elicited during three different parts, which together provided around ninety minutes of speech from every pair of speakers. While Parts 1 and 2 did not require participants to perform any specific task, in Part 3 participants negotiated a common answer to general questions about society. Comparisons with the ESTER corpus of journalistic speech show that the two corpora contain speech of considerably different registers. A number of indicators of casualness, including swear words, casual words, *verlan*, disfluencies and word repetitions, are more frequent in the NCCFr than in the ESTER corpus, while the use of double negation, an indicator of formal speech, is less frequent. In general, these estimates of casualness are constant through the three parts of the recording sessions and across speakers. Based on these facts, we conclude that our corpus is a rich resource of highly casual speech, and that it can be effectively exploited by researchers in language science and technology.

Key words: corpus; casual speech; French.

1. Introduction

French is one of the best documented languages in the world. Accordingly, researchers interested in spoken French have a choice among several speech corpora for their studies (e.g. ESTER (Galliano et al., 2005), PFC

ACCEPTED MANUSCRIPT

(Durand et al., 2005), see `http://catalog.elra.info/` for more). However, no existing corpus of French contains the large amounts of casual speech necessary for detailed research on the characteristics of this register, including inter- and intra-speaker variability. This article describes a new corpus of European French that fills this gap.

The specific characteristics of a given corpus present advantages and disadvantages depending on the researcher's goals. For instance, the ESTER corpus, with around 90 hours of journalistic recordings, mainly comprises prepared speech, either planned or read, from several European and North-African French-speaking radio stations. It is a valuable source for researchers interested in journalistic speech covering a broad range of topics with a huge lexical variety, produced by a large population of professional and occasionally intervening speakers in various audio conditions.

The PFC (Phonologie du Français Contemporain) corpus, which is still under development, contains recordings of speakers from diverse geographic and social backgrounds in the French-speaking world. In the future, it will consist of a total of several hundred hours of speech of different registers, thus providing a reference corpus for linguists who are interested in social and regional language variation. Every informant in the PFC corpus contributes an average of 20 minutes of speech, including both read speech and conversations with the interviewer. The conversations differ in their degree of casualness, as the interviewers and informants are not all on familiar terms. So far, only a relatively small percentage of these data has been orthographically transcribed. The recordings, which are field data collections, are of variable acoustic quality and hence not always appropriate for detailed acoustic analysis.

The motivation behind the creation of the Nijmegen Corpus of Casual French (NCCFr from now on) was to provide large amounts of high-quality recordings of casual speech suitable for phonetic analysis. The uniqueness of our corpus can be characterized as follows:

- It contains large amounts of casual speech. All of the recorded speech has been orthographically transcribed.
- It contains high-quality recordings captured with head-mounted microphones in a sound-attenuated room.
- It contains speech from 46 speakers sharing the same geographic and educational background. This allows researchers to study inter-speaker

ACCEPTED MANUSCRIPT

variation in a corpus controlled in terms of regional and social variation.

- It contains large amounts of data for every speaker (around 90 minutes of recorded conversation for every pair of speakers). This allows researchers to study within-speaker variability.

- It contains audio as well as video data, which can be used to study facial and body gestures during verbal communication.

Information about how to obtain a copy of the corpus can be found online at `http://mirjamernestus.ruhosting.nl/Ernestus/NCCFr`.

The present article provides a detailed description of the creation and characteristics of our corpus (Sections 2 and 3). In Section 4 we provide evidence of the casual register of the speech contained in the corpus by comparing the NCCFr and the ESTER corpus in terms of several uncontroversial indicators of casualness (e.g. lexical items sensitive to register, double negation, disfluencies). We also assess the variability of these indicators throughout the different parts of the recordings (see below for details) and across speakers.

2. Corpus creation

2.1. Participants

The corpus creation was initiated in November 2007. Twenty-three confederates were recruited at the University of Paris 3 Sorbonne Nouvelle, either by e-mail or personally. These confederates were briefly interviewed and asked to find two friends willing to participate in recordings of natural conversations. These friends will be referred to as *speakers* from now on. Every recording consists of a conversation among these 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 had completed the secondary education cycle in France.

- They had been raised in Central/Northern France.

ACCEPTED MANUSCRIPT

- They reported not suffering from any pathology related to speech or hearing.

In total there were 46 speakers (24 female and 22 male). Thirty-four speakers came from the Paris region. The remaining 12 came from other regions in Central and Northern France. Except for two female speakers in their fifties, all speakers were university students aged between 18 and 27. The gender, age and regional background of each speaker is provided in Table 1.

2.2. Recording set-up

The recording room was sound-attenuated and had an approximate size of 4 x 3 m. The participants sat on chairs around a table placed in the middle of the room. The confederate always sat on the south side of the table, while the speakers occupied the chairs on the north and west sides. Figure 1 shows the layout of the recording room.

For technical reasons, only two participants could be recorded. Given this limitation, we decided to dispense with the speech of the confederates, who, contrary to the speakers, were not naive about our goals and procedures. However, the confederates also wore a microphone in order to reinforce the speakers' impression that they did not have a special role in the conversation. Speakers were recorded in separate audio channels. The recording equipment consisted of an Edirol R-09 solid-state stereo recorder, Samson QV head-mounted unidirectional microphones and a stereo microphone preamplifier. Microphones were placed at an average distance of 5 cm from the left corner of the speakers' lips. The sampling rate used was 48 KHz, while quantization was set to 32 bits.

The conversations were filmed using a Canon XM2 Mini-DV video camera. The camera was placed on a tripod in a corner of the recording room as shown in Figure 1. The positioning of the camera allowed us to film the two speakers, but not the confederate. This is illustrated in Figure 2. Since awareness of being filmed could compromise the casualness of the conversations, we tried to make the speakers 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. It should also be noted that there were several other objects in the room,

Speaker	Gender	Age	Region	Speech		Speaker	Gender	Age	Region	Speech
M01L	M	23	Île-de-France	45:27		F12R	F	21	Poitou-Charentes	55:45
M01R	M	25	Île-de-France	43:23		F13L	F	19	Île-de-France	39:38
M02L	M	24	Île-de-France	55:15		F13R	F	18	Nord-Pas-de-Calais	31:49
M02R	M	24	Île-de-France	49:24		F14L	F	20	Île-de-France	28:46
F03L	F	50	Île-de-France	45:09		F14R	F	23	Île-de-France	55:33
F03R	F	40	Île-de-France	40:28		M15L	M	27	Limousin	28:24
F04L	F	25	Île-de-France	19:49		M15R	M	23	Centre	52:53
F04R	F	25	Île-de-France	40:50		F16L	F	19	Île-de-France	44:08
F05L	F	19	Île-de-France	26:39		F16R	F	21	Île-de-France	58:17
F05R	F	18	Île-de-France	28:38		M17L	M	20	Île-de-France	33:52
F06L	F	21	Île-de-France	39:48		M17R	M	20	Île-de-France	33:02
F06R	F	21	Île-de-France	49:38		M18L	M	20	Île-de-France	40:18
F07L	F	20	Île-de-France	47:25		M18R	M	20	Île-de-France	27:46
F07R	F	20	Île-de-France	40:14		M19L	M	19	Haute-Normandie	29:23
F08L	F	21	Île-de-France	40:49		M19R	M	26	Île-de-France	27:40
F08R	F	20	Île-de-France	21:29		M20L	M	22	Bourgogne	34:53
M09L	M	24	Île-de-France	29:22		M20R	M	22	Bretagne	30:23
M09R	M	24	Île-de-France	35:33		M21L	M	21	Île-de-France	37:54
F10L	F	19	Île-de-France	43:36		M21R	M	19	Bretagne	24:04
F10R	F	20	Île-de-France	41:51		M22L	M	19	Basse-Normandie	27:20
M11L	M	18	Île-de-France	46:38		M22R	M	19	Haute-Normandie	27:39
M11R	M	18	Île-de-France	46:18		M23L	M	20	Île-de-France	34:17
F12L	F	22	Île-de-France	53:56		M23R	M	23	Île-de-France	33:41

Table 1: Gender, age, regional background (region of longest residence) and total amount of recorded speech (minutes:seconds) for every speaker in the corpus. Speaker code: M = male, F = female; L = left channel, R= right channel. The number in the speaker code indicates the recording number.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

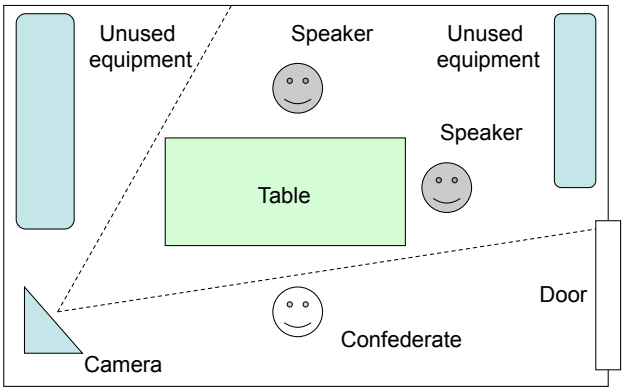


Figure 1: Layout of the recording room

ACCEPTED

including a desktop computer, several loudspeakers and other audio equipment. Our camera may have then appeared to be merely another piece of unused electronic equipment in the recording room.

2.3. Recording procedure

The recording procedure was established after a series of pilot recordings was run. During these pilot recordings, we noticed that it was difficult to obtain casual speech for long periods in the absence of any explicit task or changes in the recording setting. We also noticed that the speech recorded during the initial moments of the session was often far from casual. In order to obtain lively casual speech from our speakers for 90 minutes, we divided the session into three different parts. In the first part, in order to elicit highly casual speech right from the beginning of the recording session, the

ACCEPTED MANUSCRIPT



Figure 2: Snapshot extracted from one of the films in the corpus.

two speakers were unexpectedly left alone on the false grounds that the confederate's microphone was defective. In the second part, the confederate returned to the booth and all participants engaged in free conversation. In the third part, participants were explicitly asked to perform a communicative activity in which they had to express and negotiate their views on real issues.

We now describe the recording preparations and each of these three parts in more detail. The recordings were conducted by the first author (FT from now on). FT is not a native speaker of French, but is highly proficient in this language.

Preparations: Confederates arrived at the Institut de Linguistique et Phonétique Générales et Appliquées (ILPGA) thirty minutes earlier than their friends. At this time, FT informed the confederates that it was their task to elicit natural speech from their friends, by raising familiar topics whenever the conversation seemed to approach a dead end. In order to maximize the amount of recorded speech from the speakers, confederates were instructed to try not to monopolize the conversations. The confederates were also informed that the conversations would be filmed, and where to sit so that only the other participants would appear in the film. They were asked not to unveil any of these details to their friends until the end of the recording. Finally, they were briefly instructed about the activity planned for the third part of the recording (see below for details). Moreover, they were asked to inform their friends that the instructions for this activity were the only reason for coming to the ILPGA earlier than them. At the end of the instruction

ACCEPTED MANUSCRIPT

section, confederates were asked to wait for the other participants in the entrance hall of the ILPGA.

At the time of the appointment, FT met the three participants at the entrance hall and asked them to wait while he got the keys of the recording room. He then returned to the recording room, started the video recording, turned off the lights and locked the door. Back at the entrance 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 the confederate's seat. Once in the room, participants were asked to stay seated and not to touch their microphones or play with any other object (e.g. keys, watch) during the recording.

Part 1: After adjusting the recording volume during the first two minutes of the conversation, FT entered the recording room 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 test a new microphone. 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. In our opinion, this situation elicited very natural speech from the beginning of the recording (see Section 4 below).

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 and join their friends. The conversation then held by the three friends constituted the second part of the recordings. The topics addressed during this part were usually a follow-up to those addressed in the first part, but with the novelty of a new participant. Among the conversation topics addressed by the speakers during this part were their college exams, the ongoing strike, parties, and travel plans. No instructions were provided about the topics to be discussed during this part of the conversation.

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 (see Appendix), and then negotiate a unique answer for every question. In order to encourage them to negotiate rather than just discuss their views, we informed that they would have to write down their answers at the end of the recording session. The average duration of this part was around forty minutes.

At the end of the recording, we revealed our procedures to the participants

ACCEPTED MANUSCRIPT

and 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 and explained to them that they should only sign it if they fully agreed with its content, and that refusing to sign it would not cause them any trouble. Furthermore, we offered participants the opportunity to add restrictions to the distribution of the corpus. All of the participants signed the consent form. Two participants required that their recordings be not distributed online.

2.4. Orthographic transcription

2.4.1. Transcription protocol

The corpus was orthographically transcribed by two professional transcribers using the TRANSCRIBER software (Barras et al., 2001) following transcription guidelines developed at LIMSI (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur) in line with the French GARS conventions (Blanche-Benveniste, 1990). The transcribers were recruited on the basis of earlier transcription experience in several French and European projects with interactive spontaneous speech (e.g. the ESLO/ESLO2 project (Serpollet, 2007), the SNCF Recital project (see `http://recherche.sncf.com/uk/projets/uk_recital.html`), the ARISE project (Lamel et al, 2000)).

The speech of each speaker was orthographically transcribed in a separate annotation file using mono-channel audio streams. The audio stream was manually segmented, separately for each speaker, into small chunks of a few seconds. Most chunks contain either speech or long silent pauses, but some consist entirely of speaker noises (e.g. laughter). The transcription guidelines stated that speech in a chunk should have a clear degree of syntactic and semantic coherence and contain no long stretches of silence. In total, over 83 000 chunks were marked, with an average duration of 3.12 seconds.

Transcribers were asked to provide standard orthographic transcriptions following the French *Robert* dictionary (Le Petit Robert, 2007) wherever possible. Hence all pronunciation variants of the same word (e.g. those resulting from the addition of final schwas (Fagyal, 1998)) were annotated with the same orthographic form. However, not all speech events can be handled by a normative written language dictionary. Transcription problems arise for mispronounced words, words with an uncertain spelling (proper names, neologisms, onomatopoeia, slang...) and for unintelligible parts. The guidelines provided a series of special symbol affixes to annotate these speech events.

ACCEPTED MANUSCRIPT

Event type	Symbol	Example
Mispronunciation	* prefix	**légaliser* for [legazile]
Proper name	^^ prefix	^^*Joffroy*
Verlan	^^ prefix	^^*chelou* (for *louche*)
Standardized abbreviation	$ suffix	*fac*$ (for *faculté*)
Truncated words	( )	*re(gardes)* (for *regardes*)
Interjection	& prefix	&*ben*, &*pff*, &*euh*

Table 2: Transcription symbols.

Table 2 lists the most important ones. Notice that the same affix was used for proper names with an uncertain spelling and *verlan* words (for an explanation of the term *verlan* see Section 4.3 below). Since proper names always start with a capital letter, this convention does not lead to confusion between these two types of words.

Although the Transcriber tool proposes a list of specific noise labels, our transcribers were encouraged to fall back to a generic noise label *[b]* for all noises except for frequent and easily identifiable noises such as respiration and laughter. The label *[r]* was used for respiration noises and the label *[rire]* ('laughter') for laughter. These three labels (i.e. *[rire]*, *[r]* and *[b]*) account for 87% of all noise labels in the NCCFr corpus.

Transcribers were asked to restore common elisions and contractions to their full orthographic forms. For instance, the guidelines specified that expressions characteristic of casual speech such as *y a* 'there is' or *j'sais pas* 'I don't know' should be transcribed in their full forms as *il y a* and *je sais pas*. The reason for requiring standard full forms is that providing detailed transcriptions is very time consuming and error prone (Ernestus, 2000). Moreover, non-standard transcriptions make searching for particular lexical items difficult. However, one exception was made to this rule: the guidelines recommended that cases of obvious *ne* deletion (in the French double negations, such as *ne* ... *pas* and *ne* ... *plus*, see Section 4.4) should not be restored in the transcription. In case of doubt the *ne* particle should be transcribed.

Figure 3 shows an excerpt of a conversation with transcribed chunks of various relatively small lengths. To restore the conversation structure, the individual speaker transcription streams have been merged in the figure. The produced transcription files are in a machine readable XML mark-up

ACCEPTED MANUSCRIPT

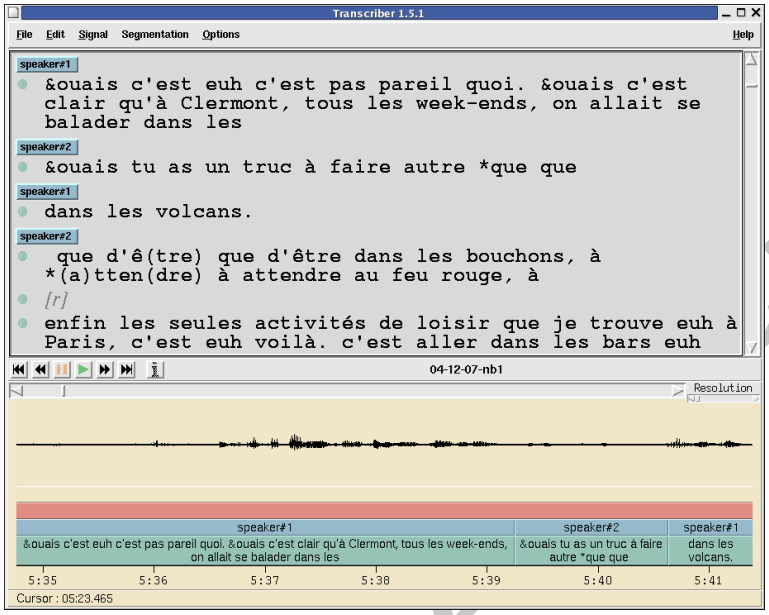


Figure 3: Manual transcription sample illustrating the segmentation and transcription steps. The two speaker streams are merged to restore the conversation structure.

language format, and are also available in Praat TextGrid format (Boersma and Weenink, 2009).

2.4.2. Transcription quality check

The quality and consistency of manual transcripts can be assessed via automatic alignment with the acoustic signal: successfully aligned parts of the corpus guarantee a good fit between the manual transcripts and the recorded speech signal. We followed this approach in order to check the quality of our transcriptions. We first created a pronunciation dictionary containing all words in the corpus and their canonical pronunciations, which were then used in an automatic alignment with the speech signal.

The corpus contains 15 919 distinct word types, including over 10% of word fragment types (truncated pronunciations). About 14 000 entries already existed in the LIMSI transcription system vocabulary (which comprises around 200 000 entries). The additional 2 000 items were checked and added to the system vocabulary with appropriate pronunciations. Among these new entries, around 1 000 correspond to word fragment types. The other entries

ACCEPTED MANUSCRIPT

include *verlan* words, interjections, onomatopeia and apocopes.

We then segmented the audio stream into words given the orthographic transcription and the pronunciation dictionary using the LIMSI recognition system (Gauvin et al., 2005). If a given transcription does not fit with the corresponding audio chunk, the alignment system will tend to reject the chunk without producing the alignment. The quality of the transcriptions can therefore be measured via chunk rejection rates. Only 41 speech chunks out of around 83 000 were rejected, corresponding to less than three minutes of speech. This strongly suggests that the orthographic transcription is of a high quality.

Transcribers often feel uncomfortable with the transcription rule prescribing the restoration of omitted and contracted words. For this reason, we also checked the extent to which the transcribers followed this rule by manually examining the two most common types of restoration: (1) *il y a* 'there is' and morphologically-related word sequences (e.g. *il y avait* 'there was'), and (2) the pronoun *tu* 'you' followed by a verb starting with an /a/ vowel (e.g. *tu as* 'you have'), in which the *tu* subject pronoun tends to be pronounced as [t], becoming a homophone with the object pronoun *t'*. Whereas for *il y a* both transcribers observed the rule in almost 100% of the cases, instances of *tu* followed by an /a/-initial verb were only restored in less than a third of the occurrences (842 out of 2 316). A possible explanation for this tendency might be that *t'* exists as an orthographically correct form for the object pronoun: *il t'a vu* ('he has seen you') is perfectly correct, but *t'as vu le film* ('you have watched the film') is not correct in written French. The manual transcripts have been updated to restore these elisions and contractions.

3. Corpus contents

The NCCFr consists of 23 recordings involving a total of 69 participants (23 confederates and 46 speakers). As explained in Section 2.2, only the speech of the two speakers was recorded. In most cases, however, the speech of the confederate was captured by the speakers' microphones and can be well interpreted from the speakers' recordings.

Table 3 shows the amount of speech contained in the corpus, both in total and averaged by recording. *Effective speech* includes all stretches of the recording containing speech by one of the two speakers, or by both at the same time (*overlapping speech*). This was calculated by adding the durations

ACCEPTED MANUSCRIPT

	Total	Average	sd	min	max
Effective speech	26h 07' 04"	1h 08' 08"	12' 25"	50' 13"	1h 30' 20"
Overlapping speech	3h 21' 50"	8' 46"	4' 44"	3' 30"	23' 46"
Non-effective-speech	10h 05' 25"	26' 19"	9' 47"	9' 54"	46' 19"
Total	36h 12' 29"	1h 34' 27"	7' 45"	1h 22' 50"	1h 52' 25"

Table 3: Amounts of *effective speech*, *overlapping speech* and *non-effective-speech* in the corpus, along with averages per recording, plus standard deviations and ranges. *Non-effective-speech* includes speech from confederates.

of all chunks in the transcriptions of each speaker containing at least one lexical item and subtracting the duration of *overlapping speech*. *Overlapping speech* was calculated by summing up the durations of stretches of conversation where both speakers spoke simultaneously. Overlapping speech involving a confederate and a speaker could not be estimated, since the speech of the former was not transcribed. *Non-effective-speech* includes stretches of conversation not containing *effective speech* by any of the two speakers. It does not only include silence, laughter and other speaker noises, but also non-overlapping speech from confederates. Overall, the corpus contains over 36 hours of recorded conversations, with over 26 hours of *effective speech*, over 3 hours of *overlapping speech* and around ten hours of *non-effective-speech* including silence and speech from confederates. The considerable amount of *overlapping speech* indicates that the corpus contains highly interactive speech (Schegloff, 2000).

Table 4 shows the total and average durations of each of the three parts of the recorded conversations, along with their average amounts of *effective speech* and *non-effective-speech*. Notice that *non-effective speech* in Part 1 refers to stretches of the conversations containing silence, laughter or other speaker noises, while *non-efective-speech* in Parts 2 and 3 also contains speech turns from confederates (remember that confederates were not recorded and only participated in Parts 2 and 3). It can be seen from this table that Parts 2 and 3 appear very similar in the percentages of effective speech that they contain, and that the inclusion of a confederate in these two parts leads to a similar decrease in *effective speech* with respect to Part 1. This was confirmed by a series of two-tailed t-tests showing that the percentage of *effective speech* in Part 1 differed significantly from those of Parts 2 ($t = 3.06, p < .005$) and 3 ($t = 3.6, p < .001$), but it did not differ between the

ACCEPTED MANUSCRIPT

	Total	Average	*Effective Speech*	*Non-effective-speech*
Part 1	7h 32' 07"	19' 39"	16' 40" (85%)	2' 58" (15%)
Part 2	12h 52' 11"	33' 34"	23' 22" (70%)	10' 13" (30%)
Part 3	15h 48' 09"	41' 13"	28' 05" (69%)	13' 08" (31%)

Table 4: Duration of each recording part in the corpus and on average per recording, along with average amounts of *effective speech* and *non-effective-speech* for each part. *Non-effective-speech* includes speech from confederates.

latter two ($t = 0.33, p = .74$).

The right-most column in Table 1 shows the total amount of recorded speech for every speaker. These amounts ranged from roughly twenty to sixty minutes, with an average of 38 minutes and 27 seconds and a standard deviation of 10 minutes.

4. Assessing casualness

In spite of our efforts to create an informal atmosphere during the recording sessions (for instance by inviting groups of friends), it is possible that speakers felt intimidated or inhibited by the awareness of being recorded. Therefore, in the absence of any proof to the contrary, the casualness of the speech contained in our corpus may be legitimately questioned. In this section we examine several indicators of spontaneity and casualness that can be extracted automatically from an orthographic transcription, and compare their values in our corpus and in the ESTER corpus of journalistic speech.

4.1. Disfluency words

We believe that genuine casual speech should be, among other things, unprepared and spontaneous. For this reason, we first quantify the incidence of disfluencies by identifying transcribed filled pauses (e.g. *euh*, *hum*, *ben*), which are known to be more frequent in spontaneous speech than in more careful and formal styles (e.g. Clark and Wasow, 1998; Shriberg, 2001; Tree, 1995; Clark, 1996). Following Jousse et al. (2008), we measured the frequencies of the most common word types used by transcribers to annotate filled pauses and hesitations, that is *ben*, *euh* and *hum*. These words will be referred to as *disfluency words* from now on. The transcription guidelines of the NCCFr and the ESTER corpus do not differ in how they specify the

ACCEPTED MANUSCRIPT

	NCCFr		ESTER	
ben	2.77	(1 292)	.26	(176)
euh	24.77	(11 546)	8.14	(5 452)
hum	9.42	(4 391)	.11	(73)
Rep. bigrams	12.94	(6 034)	3.76	(2 522)
Rep. trigrams	1.74	(815)	.49	(333)

Table 5: Frequencies of *euh*, *hum*, *ben* and repetition bigrams and trigrams in the two corpora per thousand words. Numbers within brackets indicate absolute numbers of occurrences.

annotation of filled pauses, and we have not noticed salient differences in the transcription of filled pauses between the two corpora. Therefore, we assumed that the transcribers of both corpora annotated filled pauses with the same accuracy and following similar principles. The first three lines in Table 5 suggest that the NCCFr corpus contains considerably more filled pauses and hesitations than the ESTER corpus, and thus confirm our expectations.

4.2. Word repetitions

We also counted the number of word bigrams and trigrams consisting of identical words occurring in each corpus. Following Jousse et al. (2008), we assume that word repetitions mostly result from breakdowns during online speech planning and are therefore characteristic of spontaneous speech. The word bigrams *vous vous* and *nous nous*, which form grammatical sequences (e.g. *vous vous voyez* 'you see yourself', *nous nous connaissons* 'we know each other'), were excluded from the repetition bigram count. It should be also noticed that repetitions can be used as a stylistic device to intensify the meaning of a word (e.g. *trop, trop* 'very, too much', *partout, partout, partout* 'everywhere'), or as a backchannel utterance (e.g. *oui, oui...* 'yes'). Importantly, however, the latter two types of word repetitions are also characteristic of casual speech. We therefore did not exclude these sequences from our counts. The last two rows in Table 5 show that sequences of repeated words are more frequent in our corpus than in the ESTER corpus. Again, these numbers suggest that the NCCFr contains more spontaneous speech than the ESTER corpus.

ACCEPTED MANUSCRIPT

4.3. Lexical items

An obvious way of assessing the casualness of a corpus is to check the extent to which it contains lexical items typical of casual speech. In order to do this, we examined the frequency of occurrence of swear words and *verlan* (see below for details), and also compared the use of informal and formal words with a similar meaning. We determined which casual words and swear words would be considered for analysis by asking four native speakers of French to provide two lists. The first list should contain a subjective choice of the ten most common French swear words, while the second should consist of formal and informal content words having similar meanings (e.g. *chose / truc* 'thing'). The lists of swear words were very similar, as six terms were present in all of them. On the other hand, only two pairs of formal and informal words were present in all of the second lists.

From the first lists, we selected for analysis those swear words that occurred at least ten times in either the NCCFr or the ESTER corpus. The threshold was set at ten so that interpretable comparisons could be made (comparisons of very low frequencies, say three and one, would have been hard to interpret). This threshold also allowed for a reasonable number of comparisons between the corpora. Table 6 shows the frequency of occurrence for each of these swear words in both corpora. We were surprised to find out that swear words were highly frequent in our corpus (e.g. on average, *putain* occurs roughly once every six minutes of conversation). In our view, such a frequent usage of swear words constitutes strong evidence of the casual speech register of our recordings (Eggins and Slade, 1997).

From the second lists, we only retained those pairs of which each member appeared at least ten times in one of the two corpora. When two pairs shared the same formal word, they were reorganized into a triplet (e.g. formal: *garçon*; informal: *mec*, *gars* 'lad'). We added two pairs of function words (i.e. *cela / ça* and *oui / ouais*) which in our opinion are very good indicators of register as well. Notice that our subjects had been asked to provide pairs of content words, and had therefore not mentioned any of these two function words. Table 7 shows the frequencies of occurrence of these words in both corpora. This table shows that all casual words are more frequent in the NCCFr than in the ESTER corpus. Moreover, some casual words in our corpus are more frequent than their more formal synonyms (e.g. *ça*, *ouais*, *truc* and *mec* occur more often than *cela*, *oui*, *chose* and *garçon*). So far, these facts lead us to conclude that our speakers did not generally aim at a formal register of speech during the conversations.

ACCEPTED MANUSCRIPT

	NCCFr		ESTER	
chier	.23	(110)	.00	(0)
con	.21	(102)	.00	(2)
cul	.06	(31)	.00	(0)
merde	.32	(152)	.00	(1)
putain	.79	(370)	.00	(0)

Table 6: Frequencies of occurrence per thousand words for five swear words in the NCCFr and the ESTER corpus. Numbers within brackets indicate absolute numbers of occurrences.

We also checked the usage of *verlan* in our corpus. *Verlan* is a language game typically consisting in the inversion of segments and syllables in a word, often accompanied by other changes, affecting for instance the quality of vowels. The name *verlan* /vɛʁlɑ̃/ itself is an example of such inversion, as it comes from *l'envers* /lɑ̃vɛʁ/ 'the inverse'. Importantly for our purposes, the use of *verlan* can be used as an indicator of casualness, as it is common in slang and youth language (Valdman, 2000). *Verlan* word types used in the NCCFr were identified on the basis of the prefix ^^ in the orthographic transcriptions (see Section 2.4.1). There was a total of 14 word types and 232 tokens of *verlan* words. The most frequent ones ($n > 10$) are listed in Table 8 with their number of occurrences. It should be noticed that none of these words appeared in the ESTER corpus. The occurrence of *verlan* in the NCCFr corpus constitutes further evidence that it contains highly casual speech.

4.4. Double negation

Negation in French requires the use of two grammatical particles, the first of which must be *ne* (or its contracted form *n'* before a vowel). For instance, in the utterance *Je ne veux pas dormir* 'I don't want to sleep', the negation particle *pas* appears after the verb *veux*, while the negative particle *ne* precedes it. In the same way, the word *ne* occurs along other negative particles such as *rien* 'nothing', *jamais* 'never' or *aucun* 'any'. Importantly for our purposes, casual French is characterized by the frequent elision of the particle *ne* (Coveney, 1996; Armstrong and Smith, 2002). For instance, in informal settings *Je veux pas* 'I don't want' is often heard instead of *Je ne veux pas*.

ACCEPTED MANUSCRIPT

Word	NCCFr		ESTER		Gloss
ami(s)	.13	(65)	.14	(94)	friend(s)
pote(s)	.16	(87)	.00	(1)	
argent	.21	(98)	.11	(74)	money
tune(s)	.05	(29)	.00	(0)	
cela	.01	(6)	.55	(369)	that (Pron.)
ça	17.75	(8 276)	1.72	(1 152)	
chose(s)	1.25	(587)	.53	(358)	thing(s)
truc(s)	3.01	(1 400)	.00	(1)	
fille	1.13	(531)	.07	(51)	girl(s)
nana(s)	.11	(52)	.00	(0)	
fou	.15	(71)	.02	(18)	crazy
dingue	.08	(39)	.00	(2)	
garçon(s)	.58	(271)	.02	(20)	lad(s)
gars	.48	(226)	.00	(2)	
mec(s)	.67	(315)	.00	(2)	
livre(s)	.05	(27)	.14	(97)	book(s) (N)
bouquin(s)	.01	(48)	.00	(2)	
mange(r)	.17	(81)	.03	(24)	eat(s) (V)
bouffe(r)	.08	(41)	.00	(3)	
oui	6.32	(2 949)	.83	(558)	yes
ouais	17.89	(8 343)	.02	(15)	
travail	.19	(90)	.35	(235)	work (N)
boulot	.08	(38)	.00	(4)	
très	1.33	(622)	1.64	(1 099)	very
vachement	.30	(141)	.00	(0)	

Table 7: Frequencies per thousand words for casual (in bold) words and their standard variants in the two corpora. Numbers in brackets indicate absolute numbers of occurrences. (V = verb; N = noun; Pron. = pronoun)

ACCEPTED MANUSCRIPT

Verlan	Frequency	Standard form
ouf(s)	80	*fou* 'crazy'
meuf(s)	63	*femme(s)* 'woman/women'
relou	32	*lourd* 'heavy-going'
chelou	14	*louche* 'dodgy'
vénère(s)	11	*énervé(s)* 'angry'
rebeu	10	*arabe* 'arab'

Table 8: Frequent *verlan* words and their numbers of occurrences in the NCCFr.

We investigated how often negation occurred in both corpora without the first element *ne*. Our goal was to identify rough differences in the use of double negation between the two corpora, rather than make our estimates of double negation as accurate as possible. Therefore, instead of checking every instance of double negation manually, we automatically extracted the frequency of *ne* in each corpus and compared these with the automatically extracted frequencies of negation particles *pas* 'not', *rien* 'nothing', *jamais* 'never' and *aucun(e)* 'any'. Negation particles whose orthographical form may also occur with other meanings (e.g. *personne* 'person' and 'nobody'; *que* 'that' and 'only') were not examined. An exception was made for *pas*, since it occurs far more often as a negative particle ('not') than as a noun ('step').

Table 9 shows the frequencies of occurrence of *ne* and of other negation particles in both the NCCFr and the ESTER corpus. An estimate of the percentage of double negation usage was computed by dividing the number of *ne* occurrences by the total number of occurrences of the other negation particles. As expected, double negation turned out to be very infrequent in our corpus (6.7%), suggesting that the register of the recorded conversations was highly casual and informal. On the other hand, the ESTER corpus exhibits a high rate of double negation (89%), confirming that a more formal register was used in the journalistic speech materials.

4.5. Homogeneity across parts

Tables 4-8 show that the NCCFr contains highly casual speech in spite of the fact that speakers were conscious of being recorded. Since the recordings consisted of three different parts, we investigated whether these parts differed in their degree of casualness. It might be expected, for instance, that

ACCEPTED MANUSCRIPT

	NCCFr		ESTER	
ne/n'	1.32	(676)	3.10	(3 664)
pas	19.31	(9 001)	4.55	(3 503)
rien	1.19	(557)	.28	(193)
jamais	.68	(321)	.26	(179)
aucun(e)	.21	(97)	.35	(239)
Double Neg. %	6.7		89.0	

Table 9: Frequencies of negation particles in the two corpora per thousand words, and estimated percentage of double negation (Double Neg. %). Numbers within brackets indicate absolute numbers of occurrences.

Part 3 contained less casual speech, since it involved discussing a number of prescribed topics. We therefore examined the distribution of lexical and disfluency indicators across the different parts.

In previous subsections in which we compared the NCCFr and the ESTER corpus, we examined indicators that occurred at least ten times in one of the two corpora. The same restriction cannot be applied to a comparison of the three parts in our recordings, since frequencies slightly above ten are too low to obtain interpretable differences in this case (e.g. an indicator with four occurrences in Part 1, two occurrences in Part 2 and six occurrences in Part 3 does not provide information about whether the three parts are different). We therefore decided to investigate only those indicators that appeared in our corpus at least 100 times (*chier*, *con*, *merde*, *putain*, *cela / ça*, *garçon(s) / gars / mec(s)*, *oui / ouais*, *très / vachement*, *ben*, *euh*, *hum*, word repetitions). The percentage of double negation, which could be reasonably well estimated for each part, was also included in this comparison.

Table 10 shows our findings. The usage of swear words and casual words does not exhibit significant differences across parts. The only exception perhaps is the word *garçon*, which was over twenty times more frequent in Part 3 than in Parts 1 and 2 combined. This increase may be explained by the fact that one of the questions included in the activity performed during Part 3 explicitly mentioned the word *garçon* (*Pourquoi les garçons et les filles ne sont-ils pas éduqués de la même manière?* 'Why aren't boys and girls raised in the same way?'). As in the case of swear words and casual words, word repetitions and double negation appear to be equally distributed across the three parts.

ACCEPTED MANUSCRIPT

We tested for systematic differences in the frequency of casualness indicators across the three parts by fitting a mixed-effects linear model with log normalized frequency as the predicted variable, recording part as predictor and casualness indicator (e.g. *ça*, *truc(s)*, *chier*, *euh*) as random factor. Since double negation was estimated as a percentage, it was not included in the analysis. From the pairs and triplets of formal and informal words, only informal words were retained for analysis. No statistical effect of recording part on log normalized frequency was identified ($F(2, 36) = 0.79, p > .1$), suggesting that parts did not differ systematically in their degree of casualness.

4.6. Homogeneity across speakers

We finally assessed the distribution of indicators of casual speech across speakers. Our goal was to check if the casual characteristics of the NCCFr revealed by our previous analyses were due only to a small group of speakers. Figures 4 and 5 show kernel density plots[1] of within-speaker estimates for every indicator of casualness.

The top left panel of Figure 4 shows that a few speakers used more swear words than the rest, but overall the distribution of swear word frequencies is skewed only very slightly. Importantly, only five out of the 46 speakers did not pronounce any of the five swear words we selected for analysis.

Casual word use (use of the casual members of casual/non-casual word pairs) was estimated following a procedure different from the one used in the previous section. This time we added up the total number of tokens of casual and formal content words listed in Table 7 pronounced by each speaker, and then calculated the percentage of casual words over this total. The reason for doing this was that, in order to plot the data, we needed a single score for each speaker, rather than multiple scores corresponding to different formal/informal pairs and triplets. Function words (i.e. *oui* / *ouais* and *cela* / *ça*) were considered separately from other casual words and are

[1]Kernel density plots display the estimated probability density function (y-axis) of a continuous random variable (x-axis), and have a purpose similar to that of histograms. However, whereas histograms group observations into a discrete number of bins, kernel density plots provide a continuous estimate of the distribution of a variable. The kernel density plots shown here were computed using the *density* function in the statistical software R with default parameters. For further details, see the R manual (R Development Core Team, 2008) and Sarkar (2008).

ACCEPTED MANUSCRIPT

	Part 1		Part 2		Part 3	
chier	.10	(13)	.14	(24)	.35	(73)
con	.26	(32)	.17	(30)	.19	(40)
merde	.38	(47)	.26	(45)	.28	(60)
putain	1.21	(151)	.72	(124)	.45	(95)
cela	.00	(1)	.00	(2)	.00	(3)
ça	15.87	(1 984)	15.19	(2 633)	17.36	(3 655)
chose(s)	.99	(124)	.91	(157)	1.45	(306)
truc(s)	3.58	(448)	2.76	(479)	2.25	(473)
garçon(s)	.10	(12)	.00	(2)	1.22	(257)
gars	.30	(37)	.64	(111)	.37	(78)
mec(s)	.46	(57)	.70	(121)	.65	(137)
oui	6.44	(805)	4.91	(851)	6.14	(1 293)
ouais	18.32	(2 290)	15.66	(2 715)	15.85	(3 338)
très	1.15	(144)	1.49	(259)	1.04	(219)
vachement	.46	(58)	.20	(35)	.23	(48)
ben	2.07	(259)	2.47	(428)	2.87	(605)
euh	24.11	(3 014)	22.02	(3 817)	22.38	(4 713)
hum	1.05	(131)	.76	(132)	.65	(136)
Rep. bigrams	11.29	(1 412)	11.99	(2 080)	12.07	(2 542)
Rep. trigrams	1.43	(179)	1.58	(275)	1.71	(361)
Double negation	4.9%		6.5%		7.9	%

Table 10: Frequencies of casualness indicators per thousand words within each recording part (lexical items, disfluency words, word repetitions) and percentage of double negation. Numbers within brackets indicate absolute numbers of occurrences. Double negation was estimated as explained in Section 4.2. Only lexical items that appeared at least 100 times in the NCCFr were included in this table. Rep. stands for repetition.

ACCEPTED MANUSCRIPT

not included in the figure, since they were overwhelmingly more frequent than content words and would have had too big an impact on the measure. Casual word use ranged from 0% to 92%. Only 8 speakers did not pronounce any of the casual words considered for analysis and therefore scored very low on this indicator of casualness. With respect to the function words, the word *ça* was used by all speakers, while the few occurrences ($n = 6$) of the more formal variant *cela* were shared by three speakers. The word *ouais* showed more variability, with 32 speakers showing a use between 30% to 95%, with a mean of 69.8%, and 14 speakers not using *ouais* at all. Interestingly, these 14 speakers used *oui* as often as the other participants. *Verlan* was used by 60% of the speakers.

The other indicators also showed that most speakers used casual speech. Double negation use was generally low across speakers, as expected from our previous analyses, with only a small number of significantly deviant speakers: three speakers displayed double negation rates between 15% and 30%, and two showed surprisingly high rates (38.9% and 55.8%). Furthermore, all speakers exhibited at least five repetition bigrams per thousand words, and disfluency words were used by all except two speakers.

We finally checked whether speakers with low scores for a specific indicator of casualness also exhibited low scores for other indicators. This appeared to be the case only for one speaker who pronounced zero casual words, two swear words and displayed a high percentage of double negations (55.88%). Notice, however, that this speaker's double negation percentage was still much lower than that found in the ESTER corpus (89%). All other speakers who happened to display a low score for one indicator did not have particularly low scores for the other indicators. We therefore conclude that, in spite of individual differences in terms of specific indicators, the vast majority of speakers produced highly casual speech.

5. Discussion

In the previous sections we have described a new speech corpus, the Nijmegen Corpus of Casual French. The corpus contains a total of over 36 hours of orthographically-transcribed recordings involving 23 pairs of speakers with similar social and geographical backgrounds. These numbers should make it possible to model and study in detail the characteristics of spontaneous speech and inter- and intra-speaker variation. Our corpus can also be

ACCEPTED MANUSCRIPT

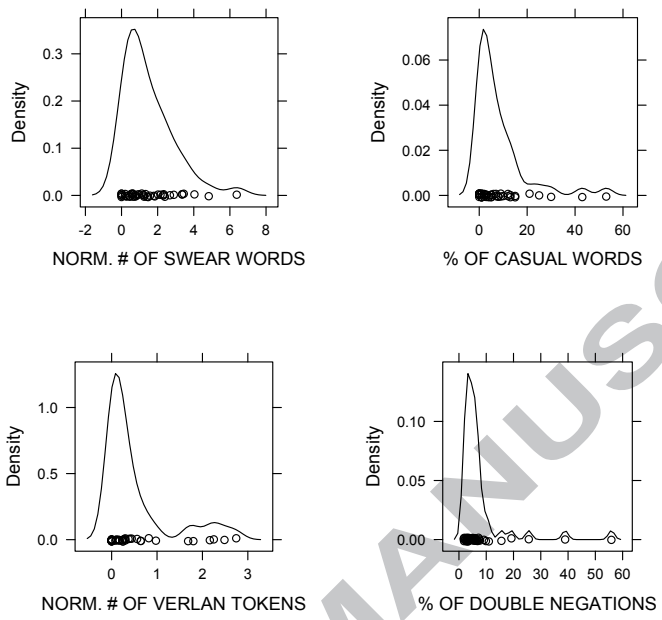


Figure 4: Kernel density plots of within-speaker frequencies of swear words (*putain*, *merde*, *chier*, *con* and *cul*) and *verlan* (normalized per thousand words), and percentages of casual content words and of double negation. Circles represent individual speakers. Except for double negation, higher values indicate higher degrees of casualness. NORM. # stands for normalized number

used to study gender differences, since gender was explicitly controlled for in our selection of speakers.

Our comparison of the NCCFr and the ESTER corpus of journalistic speech in terms of several indicators of casualness shows that our new corpus contains speech of a more casual nature. The high frequencies of swear words, casual words, *verlan*, disfluency words and word repetitions along with the low usage of double negation suggests that speakers generally aimed at a casual speech register in spite of the awareness of being recorded. The analyses in Sections 4.5 and 4.6 further suggest that this casual register was present throughout the different parts of each recording and in all speakers (excepting perhaps one male speaker). The NCCFr can therefore be used as a resource to investigate all sorts of linguistic phenomena related to casual speech, such as speech reduction (Ernestus, 2000; Johnson, 2004), disfluencies (Clark and Wasow, 1998), or the prosodic and syntactic characteristics of

ACCEPTED MANUSCRIPT

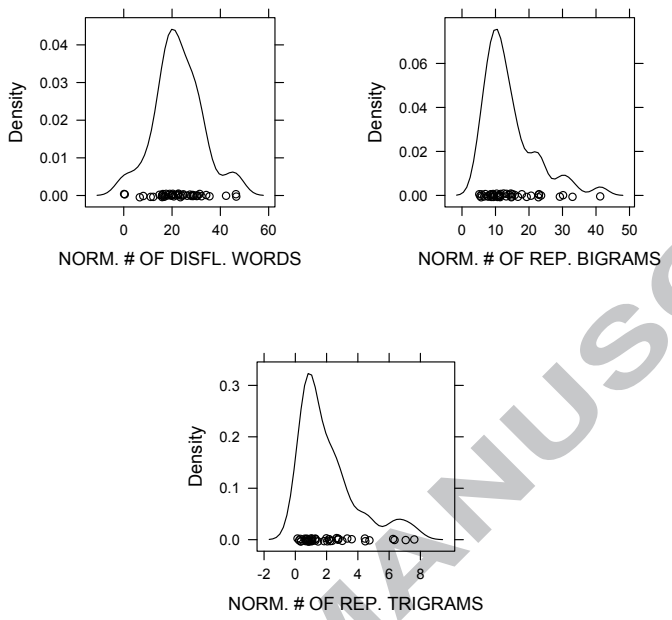


Figure 5: Kernel density plots of within-speaker frequencies of disfluency words (*euh*, *hum*, *ben*) and repetition bigrams and trigrams (normalized per thousand words). Circles represent individual speakers. NORM. # stands for normalized number

unprepared speech, among many other possible topics.

Every recording session was divided into three parts so that natural speech could be for long periods of time. A welcome consequence of this division is that specific parts of the corpus can be used to study specific phenomena. For instance, Part 1, in which speakers were left alone unaware of being recorded, is a good resource for researchers interested in talker interaction, turn-taking and conversation analysis in general (e.g. Local, 2003, 2007; Plug, 2005). Parts 2 and 3 can also be used for the same purposes, but the presence of a confederate, whose speech was not directly recorded, may complicate the study of these subjects. Part 3, in which participants were asked to choose and discuss specific topics, can be used to study argumentation and strategies used by speakers to convince their interlocutors. It can also be used to study the phonetics of specific content words, since many groups of speakers produced the same content words while discussing the same questions during Part 3. The description of the corpus provided in this article should allow

ACCEPTED MANUSCRIPT

researchers to judge which part best suits their purposes.

Finally, we hope that the corpus will be of use for researchers in different fields of speech technology. For instance, given the challenge that spontaneous speech presents to ASR systems (Moore, 2003, 2005), annotated resources such as the NCCFr may help to improve current technology.

In conclusion, the Nijmegen Corpus of Casual French is a rich source of high-quality speech data that will help researchers to study spontaneous speech from many perspectives.

6. Acknowledgments

Our thanks to Cécile Fougeron, Coralie Vincent, Christine Meunier, Ton Wempe, the staff at ILPGA and the participants for their help during the recording of the corpus in France. We also want to thank Lou Boves and Christopher Stewart for helpful comments and discussion. This work was funded by a European Young Investigator Award to the third author. It was presented at the 6^{th} *Journées d'Études Linguistiques* of Nantes University in June 2009.

References

Adda-Decker, M., Barras, C., Adda, G., Paroubek, P., Boula de Mareuil, P., and Habert, B. (2008). Annotation and analysis of overlapping speech in political interviews. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*.

Armstrong, N. and Smith, A. (2002). The influence of linguistic and social factors on the recent decline of French *ne*. *Journal of French Language Studies*, 12(01):23–41.

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.

Blanche-Benveniste, C. (1990). *Le français parlé, études grammaticales*. Éditions du CNRS, Paris.

Boersma, P. and Weenink, D. (2009). Praat: doing phonetics by computer (version 5.1.18) [Computer program]. Retrieved October 9, 2009, from http://www.praat.org/.

ACCEPTED MANUSCRIPT

Clark, H. (1996). *Using Language*. Cambridge University Press.

Clark, H. and Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology*, 37:201–242.

Coveney, A. B. (1996). *Variability in Spoken French. A Sociolingistic Study of Interrogation and Negation*. Elm Bank, Exeter.

Durand, J., Laks, B., and Lyche, C. (2005). Un corpus numérisé pour la phonologie du français. In Williams, G., editor, *La linguistique de corpus*, pages 205–217. Presses Universitaires de Rennes, Rennes.

Eggins, S. and Slade, D. (1997). *Analysing Casual Conversation*. Cassell.

Ernestus, M. (2000). *Voice assimilation and segment reduction in Dutch: A corpus-based study of the phonology-phonetics interface*. LOT, Utrecht, The Netherlands.

Fagyal, Z. (1998). Le retour du e final en français parisien: changement phonétique conditionné par la prosodie. In Englebert, A., Pierrard, M., Rosier, L., and van Raemdonck, D., editors, *Vivacité et diversité de la variation linguistique*, volume 3, pages 151–160. Max Niemeyer Verlag.

Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F., and Gravier, J. (2005). ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast Newshase II Evaluation Campaign for the Rich Transcription of French Broadcast News. *Proc. Interspeech 2005*, pages 2453–2456.

Gauvain, J.-L., Adda, G., Adda-Decker, M., Allauzen, A., Gendner, V., Lamel, L., and Schwenk, H. (2005). Where are we in transcribing French broadcast news? *Proc. Interspeech 2005*.

Johnson, K. (2004). Massive reduction in conversational American English. In Yoneyama, K. and Maekawa, K., editors, *Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium*, pages 29–54, Tokyo, Japan. The National International Institute for Japanese Language.

Jousse, V., Estève, Y., Béchet, F., Bazillon, T., and Linares, G. (2008). Caractérisation et détection de parole spontanée dans de larges collections de documents audio. JEP 2008, Avignon.

ACCEPTED MANUSCRIPT

Lamel, L., Rosset, S., Gauvain, J.L., Bennacef, S., Garnier-Rizet, M. and Prouts, B. (2000). The LIMSI ARISE System. *Speech Communication*, 31(4):339-354.

Le Nouveau Petit Robert 2008 Grand Format (2007). Editions Le Robert.

Local, J. (2003). Variable domains and variable relevance: interpreting phonetic exponents. *Journal of Phonetics*, 31:321–339.

Local, J. (2007). Phonetic detail and the organisation of talk-in-interaction. In *Proceedings of the XVIth International Congress of Phonetic Sciences. Saarbruecken, Germany: 16th ICPhS Organizing Committee.*

Moore, R. K. (2003). A comparison of the data requirements of automatic speech recognition systems and human listeners. In *Proceedings of Eurospeech-2003*, pages 2581–2584, Geneva, Switzerland.

Moore, R. K. (2005). Cognitive informatics: The future of spoken language processing? In *Proceedings of the 10th International Conference on Speech and Computer*, Patras, Greece.

Plug, L. (2005). From words to actions: the phonetics of eigenlijk in two communicative contexts. *Phonetica*, 62:131–145.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, http://www.R-project.org.

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. Springer.

Schegloff, E. A. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, 29(1):1–63.

Serpollet, N., Bergounioux G., Chesneau A., Walter R. (2007). A Large Reference Corpus for Spoken French: ESLO 1 and 2 and Its Variations. *Proceedings from Corpus Linguistics Conference Series, University of Birmingham.*

Shriberg, E. (2001). To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(1):153–169.

ACCEPTED MANUSCRIPT

Tree, J. E. F. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, 34:709–738.

Valdman, A. (2000). La langue des faubourgs et des banlieues : de l'argot au français populaire. *The French Review*, 73(6):1179–1192.

A. Appendix: Activity sheet (English translation of French original)

Now you will answer at least five from the following questions:

- What do you think about Nicolas Sarkozy's divorce and the way it has been dealt with by the media?
- In your opinion, why did Ségolène Royal lose the presidential election?
- What do you think about applying affirmative action in the government and in the workplace?
- What do you think about the smoking ban in public spaces (restaurants, bars, trains, planes)?
- What do you think about the legalization of soft drugs?
- Why aren't boys and girls educated in the same way?
- Do you think that Al Gore deserves his Peace Nobel Prize?
- What do you think about strikes in France?
- What do you think about special regimes?
- How would you improve the higher education system?

For every question, you will try to negotiate a common answer. Once the recording has finished, one of you will write down your common answers about each of the chosen questions. You will therefore need to clearly determine your common answers as well as any point for which an agreement was not possible.