Published in: Calzolari, Nicoletta; Choukri, Khalid; Thierry, Declerck; Loftsson, Hrafn; Maegaard, Bente; Mariani, Joseph et al. (Hg.) (2014): Ninth International Conference on Language Resources and Evaluation (LREC'14. May 26-31, 2014, Proceedings. Reykjavik, Iceland, pp. 2648-2652

Aix Map Task corpus: The French multimodal corpus of task-oriented dialogue

Jan Gorisch[1], Corine Astésano[2], Ellen Gurman Bard[3], Brigitte Bigi[1], Laurent Prévot[1]

[1] Aix-Marseille Université, CNRS, Laboratoire Parole et Langage, France
[2] U.R.I Octogone-Lordat (E.A. 4156), Université de Toulouse, UTM, Toulouse, France
[3] University of Edinburgh, PPLS and HCRC, United Kingdom
jan.gorisch@lpl-aix.fr corine.astesano@univ-tlse2.fr ellen@ling.ed.ac.uk brigitte.bigi@lpl-aix.fr laurent.prevot@lpl-aix.fr

Abstract

This paper introduces the Aix Map Task corpus, a corpus of audio and video recordings of task-oriented dialogues. It was modelled after the original HCRC Map Task corpus. Lexical material was designed for the analysis of speech and prosody, as described in Astésano et al. (2007). The design of the lexical material, the protocol and some basic quantitative features of the existing corpus are presented. The corpus was collected under two communicative conditions, one audio-only condition and one face-to-face condition. The recordings took place in a studio and a sound attenuated booth respectively, with head-set microphones (and in the face-to-face condition with two video cameras). The recordings have been segmented into Inter-Pausal-Units and transcribed using transcription conventions containing actual productions and canonical forms of what was said. It is made publicly available online.

Keywords: corpus, Map Task, multimodal, dialogue

1. Introduction

Leading on from pioneering work on communicative skills, the Map Task protocol was designed in Edinburgh for the HCRC Map Task corpus (Anderson et al., 1991). The usefulness of the data produced with this protocol has led many teams to create their own Map Task corpora on various languages including Italian (different varieties), Japanese or Occitan, in order to answer specific research questions on language use. However, until now no Map Task corpus was available for French.

Map Task corpora are useful in particular because the recording material can be simultaneously well controlled (in terms of lexical material, difficulty of the task, participant pairings, etc.), while allowing genuine spontaneous speech production exhibiting phenomena such as pauses, disfluencies, feedback, etc. With the collected data it is possible to target a wide range of theoretical issues. The lack of Map Task recordings in French can therefore be considered as a missing element on the way to the analysis of speech and discourse and the comparison of certain phenomena across languages. Moreover, the aim of the authors is to extend findings from read speech (Astésano et al., 2007) to speech in unscripted dialogue. The Map Task protocol in the audio-only condition serves this goal.

Although most interactions happen face-to-face, recordings of the visual part of such dialogues are still sparse. This corpus aims to provide an equivalent resource for the analysis of task-oriented dialogues with the addition of the visual modality. This makes it possible to analyse the verbal and the visual cues that are used by the participants in order to achieve their communicative goals. This consideration has led to the face-to-face condition.

This paper describes the creation of the corpus and presents its main features for the study of the audio-only and the audio-plus-visual condition of Map Task dialogues. Beside psycholinguistic studies, the corpus can also be used for the analysis of different "speech-exchange-systems" (Schegloff, 2007). Some of the speakers match participants previously recorded in free conversation, the audio-visual CID corpus (Bertrand et al., 2008). It means that it can be studied whether the same individual participant uses different communicative strategies in free conversations versus task-oriented dialogue. Other research questions that may be addressed are for example: What interactional/sequential resources and phonetic/prosodic cues do participants use when they perform feedback? What are these when they have the visual modality available as opposed to the condition without (cf. Doherty-Sneddon et al., 1997)? How does this change their feedback behaviour, e.g. do participants use more or less verbal feedback items in one or another condition; or does the choice of lexical markers change?

The corpus recordings can be classified as semi-spontaneous, task-oriented dialogues – located between the extremes of artificially elicited, controlled, read speech, e.g. Astésano et al. (2007) and naturally occurring talk-in-interaction, e.g. Bertrand et al. (2008); Kurtic et al. (2012). The present paper sets out the experimental design of the corpus (Section 2.), explains how it has been processed (Section 3.) and provides some quantitative information (Section 5.). It ends with conclusions, work in progress and plans for future research applied to this corpus (Section 6.).

2. Lexical Material and Design

2.1. Lexical Material

For both conditions, audio-only and face-to-face, the same maps were used. They contain lexical material that was selected in order to address the theoretical issues identified in earlier work on elicited speech (Astésano et al., 2007). These concern the role of Initial Accent in syntactic and prosodic structure making in French. A subset of the lexical material from the experiment on read speech was introduced in the names of map landmarks. The goal is to compare the occurrence of Initial Accents in guided dialogues with the previous results on read speech. The target words and phrases were chosen to appear on the maps,

as described in Bard et al. (2013). An example for lexical items is: "les bonimenteurs et les baratineurs fameux" (see map in Figure 1 and 2), where the likelihood of an Initial Accent on baratineurs might change with the length of the conjoined nouns or with the syntax, depending on the adjective scope (on the last noun only or on both nouns).

2.2. Experimental Design

The experimental design follows the standard rules of Map Task experiments as described in Anderson et al. (1991). The task for two participants is to collaborate in order to reproduce on the Instruction Follower's map (Figure 1) a route that is pre-printed of the Instruction Giver's version (Figure 2). Neither can see the other's map. They know that the maps describe the same features but that some details may differ. In fact, the maps differ in alternate route-critical landmarks, so that discussion of the mismatches is common. Participants are allowed to say anything necessary to accomplish their communicative goals. While the two speakers cannot see each other in the audio-only condition, any gestures spontaneously produced would not be seen by the other participant. In the face-to-face condition however, the participants could see each others' visual movements. The participants were given the instruction that the result, i.e. the success of copying the path, will be assessed afterwards.

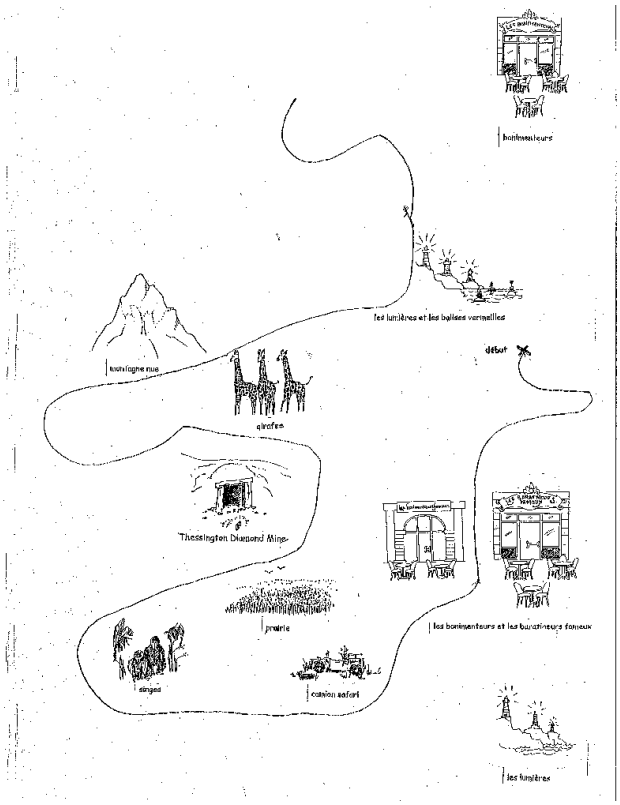

Figure 1: An example of a follower's map displayed with the route that was drawn during the experiment.

3. Data Collection

3.1. Participants

In each condition, 4 pairs of speakers performed 8 Map Tasks each (4 times as giver and 4 times as follower). In the face-to-face condition, the participants were recruited from the department where the recordings took place. They were either researchers, post-doctoral researchers or master students and knew each other. Pairs of participants were matched more or less in the sense of: researchers with researchers, students with students and according to gender (male with male and female with female). 5 sessions were recorded, involving 10 participants, 6 female and 4 male. For technical reasons, the recordings of one pair are not available, leaving 4 full sessions with 1 male and 3 female dyads.

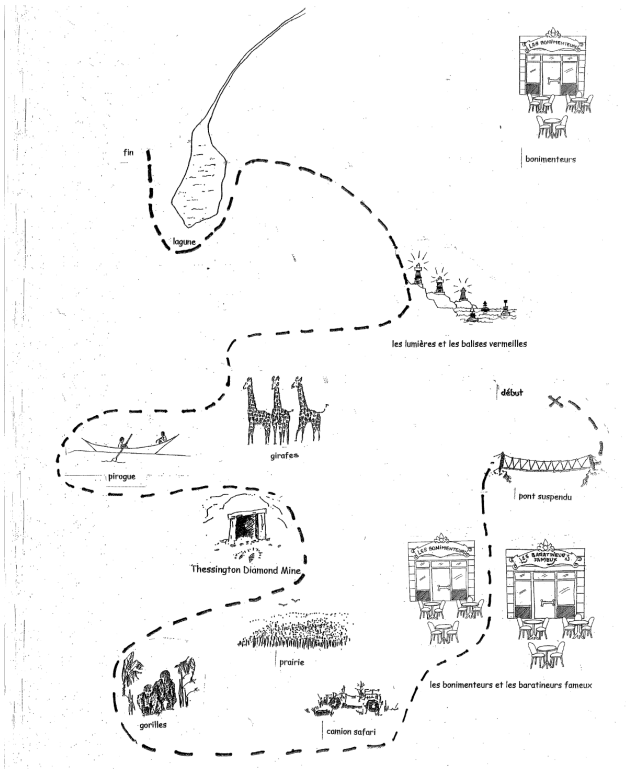

Figure 2: An example of a giver's map displayed with the pre-printed route.

3.2. Recording set-up

The audio-only condition includes one channel per speaker stereo recordings made in a studio. The recordings of the second condition took place in a sound attenuated booth. For technical reasons it was not especially necessary that the recording location was sound attenuated, as the audio was recorded on head-set microphones. However, all the equipment was already in place and sound attenuation further increased the signal to noise ratio. All participants were familiar with the room. As Figure 3 shows, members of a dyad sat face to face, each with his or her own map on an easily reached stand.

3.3. Technical aspects of the face-to-face condition

3.3.1. Digital audio recordings

The audio signal was recorded for each individual speaker via head-set microphones (AKG C520). They were connected to the video cameras and recorded on the audio channel of the cassettes. One microphone could use the phantom power of the video camera, the other was powered

Figure 3: Set-up for recording the face-to-face condition.

externally (ZOOM H4n Handy Recorder). As a backup, two additional head-set microphones were attached in parallel and transferred via XLR and an audio interface (RME Fireface UC) to a computer running Audacity[1]. Both channels are digitized with a sampling rate of 44.1 kHz and 16 bit resolution. Resampling was performed to remove an asynchrony between audio systems.

3.3.2. Video recordings

Each participant's actions were captured using separate video cameras (Figure 4). To avoid interrupting any dialogue, assettes were changed during a break after the 4th dialogue. Recordings were transferred on hard-drive and rendered under AVI format and DVSD encoding.

Figure 4: Individual camera capturing the first participant.

3.4. Segmentation, transcription and availability

The signals of the audio-only condition were transcribed in standard French orthography, using Transcriber[2] (Barras et al., 2000). The transcription includes short pauses, truncated words and hesitations and was manually cross-checked and corrected by the transcribers.

The signals of the face-to-face conditon were segmented into Inter-Pausal-Units (IPUs with 100ms minimum per speech unit and 220ms minimum per silence), using SPPAS[3]© (Bigi, 2012) and transcribed orthographically.

The remote condition is available on the SLDR (Speech and Language Data Repository)[4], as well as the audio-visual condition[5].

4. Example

An example should illustrate the kind of interaction the Map Task involves with special focus on the lexical item "ouais". Extract 1 (see next page) is taken from the audio-visual condition. Prior to this excerpt, the participants came to a point where the maps differed and a straightforward explanation of the path failed. Next to some "terrasse", the follower (F) did not have the same drawing as the giver (G). After having progressed to a common point, G comes back to that point where it is necessary to clarify what F has on his map next to the terrasse. G checks whether his assumption is correct: *'and normally you don't have any drawing on the left there' / 'et normalement tu as pas de dessin du tout à gauche là'* (l.1). The follower uses another understanding check (Kelly and Local, 1989) and repeats *'à gauche'* (l.3). It is immediately confirmed by G (*'yeah' / 'ouais'* l.4). The *'ouais'* is at least treated as a confirmation by F and interprets the point in the sequence as a place where a response to G's overall check from line 1 is necessary. F responds with a *'no' / 'non'* and adds information on the drawing that he has at that place.

The next *'ouais'* (l.9) is produced by F who takes G's guidance into account (*'you pass above' / 'tu passes au dessu-'*, that is produced in overlap, and proceeds by precising: *'but it's really in the corner at the bottom left' / 'mais c'est vraiement dans l'coin en bas à gauche'*, adding a reinforcement marker *'hein'* (*'isn't it'*). After a silence of 0.7s, G takes this information into account and F comes back to the initial point of trouble: *'however next to the terrasses above on the left of the terrasses almost glued together' / 'par contre à côté des terrasses en haut à gauche des terrasses assez collé'*. These information are taken into account by G partly in overlap, partly just afterwards. F has still the turn as the syntactic construction is not completed at that point. F indicates that the drawing from his map at that point is a *'flower bar' / 'bar fleuri'*. The two *'ouais'* employed by G serve here (l.13,14) to indicate F that he follows F or knows what F is talking about. Merely the last piece of information (*'bar fleuri'*) is not part of G's knowledge, what he indicates with *'and no I don't have that' / 'et non j'ai pas moi ça'* (l.16).

From this excerpt it is also visible that the Map Task is a very cooperative task where constant feedback is necessary in order to keep track – from the follower's as from the giver's side. Not only the giver, but also the follower takes on active behaviour. The description of his map takes equally part as the description of the giver's map. The giver gives instructions and the follower asks for instructions, as can also be seen in line 17, after the negative response from G, where F says: *'ah (.) so I have to go where there' / 'ah (.) alors il faut que j'aille où là'*.

[1] http://audacity.sourceforge.net/
[2] http://trans.sourceforge.net/
[3] http://www.lpl-aix.fr/~bigi/sppas/
[4] http://sldr.org/sldr000732
[5] http://sldr.org/sldr000875

Extract 1: Interaction between giver (G) and follower (F) from the audio-visual condition (AG-YM; minute 9:36 to 9:58). The visual movements and gaze are not included in this example transcription.

```
 1   G: .h et normalement tu as pas de dessin du tout à gauche là
 2      (0.6)
 3   F: à gauche
 4 > G: ouais
 5   F: non dans le coin en bas à gauche j'ai les lumières (0.2)
 6   F:  [et-
 7   G:  ah ben alors c'est peut-être ça[ il faut ben tu [passes au dessu-
 8   F:                                  .h              et à côté d'=
 9 > F: =la d'la::  ouais mais c'est vraiment dans l'coin en bas à gauche
10   F: hein (0.7) [.hh    [et  ] par contre à côté des terrasses en haut
11   G:             mh mh  [bon ]
12   F: à gauche [des terr [asses] assez collé
13 > G:          [o-      [ouais]
14 > G: ouais (0.2)
15   F: j'ai euh le bar fleuri (0.2)
16   G: et non j'ai pas moi ça
17   F: ah (.) [ alors il faut] que j'aille où là
18   G:        [.hh
19   G: eh ben tu tu tu remontes euh ...
```

5. Basic quantitative information

In each condition, 4 sessions of Map Task dialogues sum up to 1:50h (audio-only) and 2:18h (face-to-face) of data (see Table 1). Although the identical maps were used in both conditions, the duration is substantially longer in the face-to-face condition. This difference is also indicated by the average durations per map (see boxplots in Figure 5). Whether the time difference can be attributed to the two conditions is however speculative, as many other factors may play a role.

Table 1: Basic quantitative information per condition.

	audio-only	face-to-face
duration	6608s = 1h50	8287s = 2h18
# tokens	39792	26706

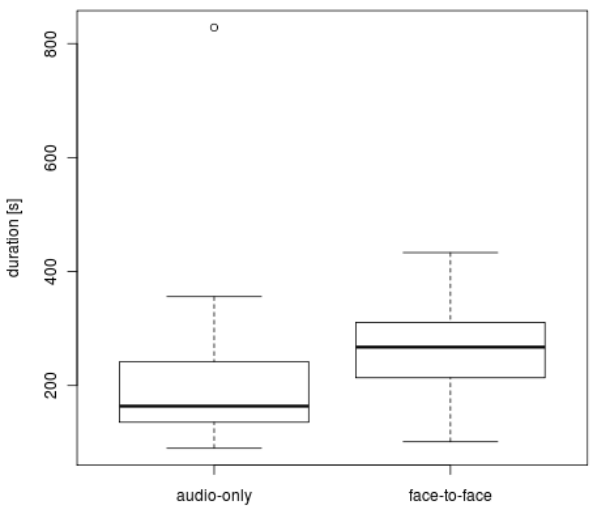

Figure 5: Duration per map.

SPPAS was used to generate initial descriptive statistics for the corpus, including separate TextGrid files for utterance, word, syllable, and phoneme segmentations.

From the word level, the relative frequencies were calculated (see Table 2). The focus is here on "feedback items", which play a role in the project CoFee (Prévot and Bertrand, 2012). It can be seen that almost 10% of all words are of the type feedback. "Ouais" is the most frequent in both conditions, followed by "mh" in the audio-only and "d'accord" in the face-to-face condition (see Figure 6). The high number of instances of feedback items: 3,669 (audio-only) and 2,271 (face-to-face) makes an extensive analysis of phonetic, prosodic, contextual and visual cues possible. The visual domain will be of special interest in the future work on this corpus, as the number of tokens vs. the duration of the dialogues diverge according to conditions: Short dialogues (audio-only) have on average more tokens than long dialogues (face-to-face). This suggests that a lot happens in the visual domain that does not happen if this resource is not available during the dialogue.

Table 2: Relative frequency of "feedback items" according to the two conditions.

	relative frequency	
token	audio-only	face-to-face
ouais	0.0269	0.0216
oui	0.0081	0.0116
voilà	0.0083	0.0071
mh	0.0207	0.0064
d'accord	0.0112	0.0176
ok	0.0059	0.0094
ah	0.0032	0.0047
non	0.0078	0.0064
total	0.0922	0.0850

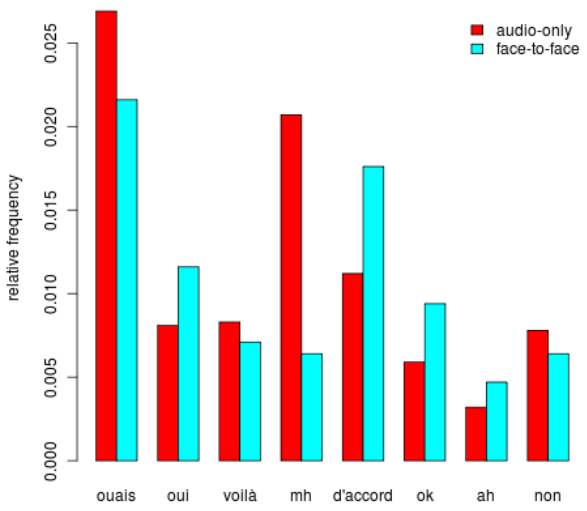

Figure 6: Relative frequency of feedback items per condition.

6. Conclusion and on-going work

The Aix Map Task corpus represents a resource for speech analysis that is both structured in terms of the selection of lexical items for intonation research and rich in spontaneity of task-oriented dialogues. It is orthographically transcribed and aligned on the word-level. The annotation of the visual movements is part of ongoing work. The integration of both modalities in the investigation of feedback items is the aim of the CoFee project where this corpus is an essential part of.

Acknowledgments

Thanks to EU Marie-Curie funding (HPMF-CT-2000-00623), the audio-only condition of the corpus was recorded and transcribed in 2002 (but never published before). With additional funding from ANR project "Conversational Feedback" (Grant Number: ANR-12-JCJC-JSH2-006-01) it has been developed for further use and supplemented by the face-to-face condition in 2013. We are grateful for the collaboration, assistance and support of colleagues the Laboratoire Parole et Langage, Bernard Teston, Thierry Legou, Noël Nguyen, Cheryl Frenck-Mestre, Mariapaola d'Imperio, Robert Espesser, Louis Seimandi, Annelise Coquillon, Ludovic Jankowski and from the University of Edinburgh, Alice Turk, Eddie Dubourg and Ziggy Campbell.

References

Anderson, A., Bader, M., Bard, E., Boyle., E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., and Weinert, R. (1991). The HCRC Map Task corpus. *Language and Speech*, 34:351–366.

Astésano, C., Bard, E., and Turk, A. (2007). Structural influences on initial accent placement in French. *Language and Speech*, 50(3):423–446.

Bard, E. G., Astésano, C., Turk, A., D'Imperio, M., Nguyen, N., Prévot, L., and Bigi, B. (2013). Aix Map-Task: A (rather) new French resource for prosodic and discourse studies. In *Proceedings of TRASP; Tools and Resources for the Analysis of Speech Prosody*, pages 15–19.

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2000). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1-2).

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., and Rauzy, S. (2008). Le CID - Corpus of Interactional Data—annotation et exploitation multimodale de parole conversationnelle. *Traitement automatique des langues (TAL)*, 49(3):105–134.

Bigi, B. (2012). SPPAS: A tool for the phonetic segmentations of speech. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1748–1755, Istanbul, Turkey.

Doherty-Sneddon, G., Anderson, A., O'Malley, C., Langton, S., Garrod, S., and Bruce, V. (1997). Face-to-face and video-mediated communication: A comparison of dialogue structure and task performance. *Journal of Experimental Psychology: Applied*, 3(2):105–125.

Kelly, J. and Local, J. (1989). *Doing Phonology: observing, recording, interpreting*. Manchester University Press, Manchester, UK.

Kurtic, E., Wells, B., Brown, G. J., Kempton, T., and Aker, A. (2012). A corpus of spontaneous multi-party conversation in Bosnian Serbo-Croatian and British English. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Prévot, L. and Bertrand, R. (2012). CoFee—toward a multidimensional analysis of conversational feedback, the case of French language. In *Proceedings of the Workshop on Feedback Behaviors in Dialog*.

Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis*. Cambridge University Press, Cambridge, UK.