Complexity versus spontaneity?: non-negotiable elements in the constitution of two interactional corpora

Adam Wilson, Mathilde Guardiola

SHS Web of Conferences **20**, 01022 (2015)
DOI: 10.1051/shsconf/20152001022

Complexity *versus* spontaneity?: non-negotiable elements in the constitution of two interactional corpora

Complexité versus spontanéité ? : éléments non-négociables dans la constitution de deux corpus interactionnels

Adam Wilson[1] and Mathilde Guardiola[1,a]

[1]*Laboratoire Parole et Langage, UMR 7309, CNRS et Aix-Marseille Université, 13100 Aix-en-Provence, France*

Résumé. Étant donnée la nature interdisciplinaire de la linguistique, il arrive fréquemment que des questions de recherche similaires soient étudiées selon différentes approches et grâce à des corpus différents. Nous traitons ici deux notions-clefs dans la conception de corpus : la spontanéité et la complexité. Nous montrons que ces dernières sont compatibles mais qu'elles entraînent de nombreux choix méthodologiques. Ces choix, ici qualifiés de non-négociables, entraînent des obligations, des contraintes et des concessions dans le déroulement du recueil de données. Nous postulons que l'existence de différents non-négociables (due à des cadres théoriques différents) offre, grâce aux corpus résultants, des éclairages complémentaires sur des objets d'étude similaires. Des notions théoriques centrales telles que la spontanéité, la complexité, la généralisabilité et la construction de données sont également discutées. Nous concluons qu'une amélioration de la description des choix ainsi faits permettrait d'augmenter le nombre et la qualité des collaborations interdisciplinaires en linguistique.

Abstract. The interdisciplinary nature of linguistics often leads to similar research questions being investigated using diverse corpora. In this paper, special attention is given to two key concerns in the corpora design: spontaneity and complexity. It is shown that spontaneity and complexity are not necessarily incompatible but often become the centre point of early methodological choices. These choices are here termed "non-negotiables" and it is demonstrated how these non-negotiables lead to obligations, constraints and concessions in the data collection process which shape the corpus. It is argued that the existence of different non-negotiables, influenced by different theoretical approaches, lead directly to the creation of different corpora. These different corpora then allow complementary lights to be shed on similar objects of study. Certain central theoretical concerns - spontaneity, complexity, generalisability and data co-construction - are also discussed. The paper concludes that an improvement in the description and diffusion of these decision processes would promote increased and improved interdisciplinary collaboration.

1 Introduction

Linguistics has traditionally been viewed as an interdisciplinary field, bringing together questions, approaches and analyses from varied domains. In the current scientific climate, this is truer than ever and, as a result, a given general research question may often be explored and addressed from many different perspectives at the same time. While (in our view), few linguists would disagree with this assessment, the diverse theoretical and disciplinary frameworks of different researchers often result in ultimately similar research questions being presented and perceived as very different. Likewise, these different ways of treating a similar question seem to be linked with working with different kinds of

[a] Auteur de correspondance : mathilde.guardiola@lpl-aix.fr

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20152001022

data. This explains the use of different corpora, created from different data and influenced by different theoretical and disciplinary frameworks, to study similar general research questions. While this may seem somewhat trivial at first sight, we argue here that taking into account this process more explicitly may allow for better understanding and communication of different approaches, helping to contribute to the climate of interdisciplinarity.

With this argument in mind, this paper presents a comparison of the conception, creation and constitution of two different interactional corpora. The aim of this paper is to illustrate how certain choices made by the researchers, partly constrained by the different approaches and the precise goals of the different studies, influence the constitution of the corpora, especially in terms of complexity and spontaneity. We argue that the link between the specific research question(s) and the type(s) of data sought is conditioned by certain choices made early in the methodological process which we will name here "non-negotiables". These "non-negotiables" will be shown to have an impact on several steps of corpus creation. We show that, despite different "non-negotiables" and resulting differences in research questions and corpus constitution, the results can be considered as complementary in terms of a general objective shared by both studies, that of exploring accommodation in human interaction.

2 Comparison of two interactional corpora

2.1 Corpus 1: MITo

MITo is a corpus of naturally-occurring interactions between tourists and tourist advisers recorded in 2014 in the Tourist Office of Marseille, France (*Office de Tourisme et des Congrès de Marseille*). It was recorded *in situ* using three discreet microphones (cf. **Figure 1**). The interactions are largely made up of exchanges of information, requests for help and commercial transactions and include interactions both in French (endolingual) and in a number of other languages (exolingual). MITo is comprised of audio recordings and annotated transcriptions of 200 interactions totalling 10 hours and 48 minutes completed by ethnographic notes taken by the researcher as well as formal and informal sociolinguistic interviews.

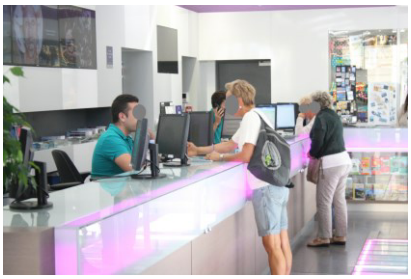

Figure 1. Example settings of MITo (left) and the CID (right)

2.2 Corpus 2: The CID

The Corpus of Interactional Data is a semi-spontaneous corpus of French conversation. It was recorded in a laboratory anechoic chamber with heavy equipment for recording the voices of participants on separate tracks. It is made up of 8 hour-long recordings, each with two same-gender participants. The participants were given the instruction to tell personal stories. One of the reasons why the CID was recorded with so much attention given to detail is due to the desire of the corpus authors to establish links between work on humour and work on gesture and prosody. These latter areas of study require high-quality video and audio recordings and this led to the CID being selected for use in a research project bringing together numerous researchers from several diverse domains.

This corpus was consequently processed in a specific manner, ensuring the interoperability of formats with different software, allowing inter-domain collaborations and facilitating future uses of the data.

2.3 Major differences between MITo and CID

Two major differences can be highlighted in the characteristics of the two corpora presented above. The first major difference between the CID and MITo concerns the spontaneity of the interactions recorded. The interactions recorded for the MITo corpus would have taken place "naturally" whether the researcher (and the equipment) was present or not. On the other hand, the interactions recorded in the CID were provoked by the experimenters.

Secondly, another difference can be found in the reasons why each corpus was recorded. MITo was designed and created to serve as a corpus for analysis in a PhD thesis (Wilson, in preparation [1]). Each detail was therefore entirely conditioned by and tailored to this particular project. The CID was conceived primarily as a resource allowing several researchers to work together. The corpus was also exploited in a PhD thesis (Guardiola, 2014 [2]) but tailoring to this doctoral research was not the main goal when constituting the corpus.

2.4 Research questions explored through these corpora

As explained in the introduction, these two corpora are used in two studies which share a similar overarching objective, that of exploring and analysing face-to-face spoken interaction. More precisely, both studies aim to describe how speakers accommodate from an interactional perspective by investigating the resources that participants use in order to adapt their production to each other within spoken interaction.

The PhD project which exploits MITo aims to analyse how participants accommodate to each other in exolingual communication in a Tourist Office. Going beyond this, the project aims to explore to what extent the use of these particular resources can be considered to contribute to the (socio)linguistic dynamics more widely associated with globalisation. The Tourist Office was thus chosen as it represents a space which acts simultaneously as a result, an example and a mirror of globalisation (Wilson, 2015 [3]).

The main goal of the PhD research which employed the CID was to describe the multimodal resources used by participants in order to converge - such as gestural, prosodic and lexical adaptation - in the co-construction of humorous sequences in story-telling. The CID was exploited in this study, but it was first conceived with a pluri-disciplinary objective, aiming to elaborate a multimodal annotation scheme allowing representation of the maximum amount of information. The existing annotations were used and enriched over the course of the PhD research.

3 Identifying obligations and constraints in corpus constitution

Despite the similar general research questions addressed in the studies presented in the previous section, they use corpora which are very different in their conception, creation and constitution. We argue here that the steps involved in determining the precise research questions lead to differences in the types of data collected and the ultimate form of the corpora. Thus, the respective theoretical approaches of each study have a direct effect on corpus constitution.

We begin from the observation that researchers are often highly reluctant to make certain concessions or compromises concerning specific parts of their research, approach and/or corpora. These "impossible" compromises will here be termed "non-negotiables" and may be defined as elements (or a single element) of corpus constitution on which the researcher refuses all concessions. These "non-negotiables" may be more or less directly determined or influenced by the approach, the methodological background, the research question(s), the research setting (and the possible constraints thereof) and/or the object of study itself. For example, as previously stated, the research project linked to MITo aims to observe the use of interactional resources in exolingual communication in a context

typical of globalisation. The "non-negotiable" in this case is then simply to collect data in such a context. The study's (and researcher's) sociolinguistic approach also influences the corpus constitution in that it favours, or, rather, renders "non-negotiable", the collection of naturally-occurring, "ecological" data (as defined by Gadet *et al*., 2012 [4]). Concerning the CID, the main objective in its construction was to create a multimodal corpus for use in various diverse research projects involving the study of gesture and prosody among other elements. The resulting "non-negotiable" is the necessity for access to as many sets of different resources used by the participants as possible through data which may be analysed phonetically, prosodically or gesturally in the finest detail possible.

These "non-negotiables" represent choices made by the researcher which take precedence above all else and are important to consider as they lead to *obligations*, *constraints* and *concessions* concerning subsequent "choices" in corpus constitution. The knock-on effects of the choice of "non-negotiable" can be seen at each of these stages in the constitution of the two corpora presented here.

In the case of MITo, the objective of study (verbal interactional resources) presents the researcher with an *obligation* to acquire spoken interactional data or, in other words, audio recording. The "non-negotiable" detailed above adds a further *obligation*; that of obtaining this data in a specific ecological setting. These obligations lead to a situational *constraint* in the choice of the research field as only a limited number of situations meet the given criteria concerning places typical of globalisation. This situational constraint then engenders certain technical *concessions* as the natural limits of the chosen situation (the Tourist Office) mean that the sound is of lower quality and difficulties concerning authorisation to record video means fewer modalities are collected than would be possible in a more controlled situation (such as a research laboratory). We can consider then that the "non-negotiable" in the constitution of this corpus, mainly concerned with preserving the nature of the situation, leads to concessions over the overall number of modalities collected (see section 4).

Concerning the CID, the "non-negotiable" regarding access to as many modalities as possible creates the *obligation* of high-quality video and audio recordings with temporal "granularity". This obligation leads to certain technical *constraints* such as the use of headset microphones and multiple cameras. These technical constraints lead to certain situational *concessions* as use of this material is only feasible in certain settings such as an anechoic chamber in laboratory conditions. In this case, the "non-negotiable" linked to preserving the number of modalities collected leads to concessions over the "naturalness" of the corpus (see section 4).

4 Are spontaneity and complexity incompatible?

In the previous section, it has been shown that, in the creation of the two corpora presented here, choices have been made between either preserving a (more) "spontaneous" (or "natural") situation or creating a more "complex" corpus through collecting more modalities. Favouring this spontaneity seems to lead to sacrifices regarding this complexity and, likewise, favouring complexity seems to lead to sacrifices regarding spontaneity. Ideally, linguists often try to maximize both spontaneity *and* complexity (as shown by the work of Mondada, 2013 [5], Mustaers & Swanenberg, 2012 [6], Tellier, 2014 [7], among others). However, as shown in the corpora presented in this paper and as demonstrated by others (see Traverso (2003 [8]), for example), this "playoff" between "complexity" and "spontaneity" seems to be a key concern in the creation of spoken corpora. These notions therefore merit further investigation in order to fully explore the choices in corpus constitution and understand whether the choice of a non-negotiable is ultimately a choice of "spontaneity" over "complexity" or vice versa.

4.1 Reconsidering (semi)spontaneity in terms of intervention and control

First of all, it is necessary to clarify exactly what is understood by the term "spontaneity". One common way to represent spontaneity is in opposition to elicitation. For exemple, Bowern (2008) contrasts "spontaneously produced speech" or "spontaneously generated data" with "elicited data" (Bowern, 2008 [9]). This dichotomy is echoed by Laurens *et al*. (2009) who oppose "ecological data",

defined as "allow[ing] for more spontaneous forms of speech", with "laboratory elicited data" (Laurens *et al*., 2011 [10]). In these definitions, spontaneity is presented as a binary notion. Corpora would thus be classified as either "spontaneous" or "elicited" ("non-spontaneous").

Following this, MITo would clearly be defined as a spontaneous corpus as the data collected is naturally occurring, or "ecological", and therefore may be classed as spontaneously generated. However, the situation regarding the CID is less clear. Stern analysis may judge the CID to be an elicited corpus due to the fact that participants are given instructions and the situation is not naturally-occurring. On the other hand, the data obtained in the CID is closer to data which may be obtained in a spontaneous corpus than data obtained through more traditional means of elicitation.

This situation has led to the qualification of the CID as a "semi-spontaneous" corpus. While open to debate, this qualification implies that spontaneity is not a binary notion. Based on the choices made in the creation of the two corpora described here, it seems pertinent to consider spontaneity in terms of two notions (among the numerous dimensions used to define spontaneity), those of "linguist intervention" and "degree of control".

"Linguist intervention" concerns the interference of the researcher in the creation of the interaction. Simply put, does the interaction exist *because of* the research? Or does the interaction exist *in spite of* the research? In the first case, the interaction is provoked by the linguist whereas, in the second, the interaction takes place without the influence of the linguist. As shown in the figure below (cf. **Figure 2**), this notion is binary: either a linguist intervenes to incite an interaction, or she doesn't. In this light, the CID is less spontaneous than MITo in the sense that the CID interactions have been initiated by the experimenters, in order to obtain speech data, whereas MITo consists of a snapshot of what would have occurred whether the researcher had been present or not.

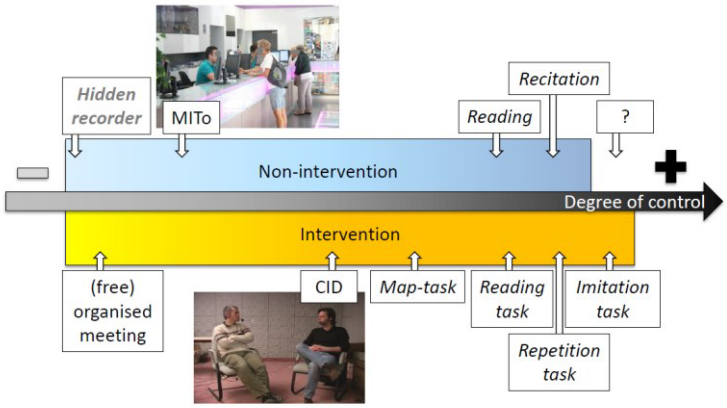

Figure 2. Figure representing spontaneity in terms of "intervention" and "control". The CID and MITo are placed on the scale with a number of sample corpora types.

"Degree of control" refers to the degree of influence on the production(s) of participants in a given corpus. This influence may take several forms, including explicit instructions given to participants, and/or may be conditioned by the task at hand in a given interaction, such as reading, imitation or repetition. This control may originate from the researcher, in the case of linguist intervention, or from other outside sources which exert an influence on a participant's production when compared with unprepared speech (such as the act of reading a prepared speech, for example). As can be seen on the scale below, we consider that the degree of control forms a continuum on which even naturally-occurring data can be graduated as any context of interaction goes some way to conditioning the production of the participants. In terms of degree of control, CID and MITo do not exhibit huge differences. While CID participants received explicit instructions, these instructions were relatively general. Likewise, the highly institutional context of MITo could suggest that participants' productions are somewhat "controlled", though this "control" is never made explicit.

This conceptualisation of spontaneity allows us to differentiate between different corpora according to different dimensions, therefore allowing us to represent MITo as more spontaneous (through its place in the non-intervention paradigm) than the CID while also allowing us to distinguish the latter from other, more controlled corpora. This depiction of spontaneity, as well as the choice of the key dimensions used to identify a "spontaneous" or "non-spontaneous" interaction, remain up for debate (not least between the two authors of this paper!) and would merit further research.

4.2 Where does corpus "complexity" come from?

In the same way that the notion of "spontaneity" in spoken corpora can be debated, so to can the concept of "complexity". Here, we base our starting definition of "complexity" on the definition of corpus "*complexe*" given in the call for papers for this issue: "*un ensemble cohérent de données "multi-modales" - associant au choix de la vidéo, du son, des textes, des traces numériques, des images, etc.*"[b] This definition seems to argue that the higher the number of modalities collected, the more a corpus may be considered complex. However, through the analysis of the two corpora presented here, the notion that there is a direct link between the number of modalities that make up a corpus and its level of complexity can be called into question. Here we argue that corpus complexity should not be purely correlated to the collection of data from different sources, nor taken as a synonym for corpus "richness".

Firstly, the definition of complexity given here focuses solely on the number of different modalities collected. However, corpus complexity could also be linked to other factors, such as the complexity of its constitution, the practical, material and ethical challenges related to developing an "ecological" corpus in certain situations or the difficulties engendered by transcribing lower-quality sound. In this case, creating a complex corpus would be about more than collecting as many modalities as possible.

Secondly, we would differentiate between the notion of corpus complexity, as it is defined above, and what could be termed "richness". Here, richness would refer to an abundance of data allowing for thorough and extensive linguistic analysis at a number of levels. It seems important to highlight the potential risk of using complexity (defined in terms of multimodality) and richness as synonyms when describing a corpus. Although multimodality may indeed increase the richness and "quality" of a corpus in some cases, it is our view that the terms "richness" and "complexity" should not be used as simple equivalents. Richness could be considered to originate from other aspects of corpus creation, such as a large(r) number of participants, a large(r) variety of interactions, the lack of direct instructions given to the participants and the knock-on effects of this on the interaction.

We would argue then that, while the number of modalities is undoubtedly important, the definition of corpus complexity should also take into account both the data richness and the other potential sources of complexity explored here. This is supported by the fact that, if we follow the definition of complexity as a collection of multimodal data, MITo would be classed as a "non-complex" corpus. However, it can be argued that the lack of control over participants can lead to unexpected behaviour, thereby increasing the potential richness of the phenomena observed. This situation is less likely in a (semi-) controlled interaction leading to a possible conclusion that, even if, in terms of modalities observed, the corpus could be termed "complex", it may not necessarily be considered "richer" in terms of phenomena observed.

Given these questions surrounding the notions of "spontaneity" and "complexity", it seems practically impossible to identify which of the corpora presented here would be the more spontaneous and/or the more complex. If we are to follow the arguments made here concerning the definitions of spontaneity and complexity, in both corpora, it is clear that *choices* have been made either to maximise spontaneity *in some way* which, in turn, seems to comprise complexity *in some way* (in the case of MITo) or the reverse (in the CID).

[b] "a coherent collection of "multimodal" data - bringing together video, sound, text, digital traces, images, etc."

This would seem to suggest that "total" spontaneity and "total" complexity are somewhat incompatible in the creation of spoken corpora. This is clearly not entirely the case as it has been shown that certain researchers (Mondada, 2013 [5], Mustaers & Swanenberg, 2012 [6], Tellier, 2014 [7]) manage to maximise both of these dimensions. Equally, questions may be asked as to what exactly is understood by the two terms. However, we would argue that, in the creation of many interactional corpora, as in the creation of ours, this playoff between spontaneity and complexity is extremely pertinent and forms the basis for most, if not all, non-negotiable choices in corpus constitution. Put another way, it would seem that the majority of non-negotiables linked to spoken corpora focus on the dichotomy: spontaneity *versus* complexity.

This conclusion may not exactly be groundbreaking but it is our belief that these considerations are often "lost" or taken for granted in the discussion and presentation of corpora, analyses and results. This is surprising given the fact that non-negotiable choices seem to define corpus constitution, bringing about inevitable differences in observations. Making these non-negotiable choices more explicit would allow researchers to better take into account their influence and therefore better understand a corpus and its accompanying analysis.

5 Complementary analysis of a complex phenomenon: repetitions

As stated in the previous section, it seems important to improve and forefront communication concerning corpus constitution, especially in terms of the corpus' "non-negotiables". This could lead to greater possibilities of "*a posteriori*" interdisciplinary exchange as it would allow more informed "crossing" of linguistic data from different studies, based on different corpora. While this question – as with many questions regarding corpus constitution – is often studied from a theoretical point of view, in this section we propose a brief application of this argument on a specific example of a pragmatic phenomenon: lexical repetition. The aim here is not to offer an exhaustive analysis of each case, but to highlight how the different analyses may be complementary.

AB: trop longtemps d'ailleurs je me rappelle pas le dernier que j'ai vu	AB: a really long time by the way i don't remember the last one i saw
CM: ah bien quand même tu dois y aller oh oui tiens ça me fait penser que j'ai plus de	CM: ah well really you must go oh yes look that reminds me that i haven't got any more
CM: euh tu en as pas racheté par exemple des des tickets [César Variété]	CM: euh you haven't bought some more for example some tickets for [César Variété]
AB: [bien moi j'en] ai encore puisque j'ai [quasiment] euh ouais	AB: [well I've got] some left because I [hardly] er yeah
CM: [tu en as encore]	CM: [you've got some left]
AB: je suis quasiment pas allée au ciné	AB: I've hardly been to the cinema

Figure 3. Example of confirmation request repetition in the CID in the original French (left) and a translation (right)

The example above (cf. **Figure 3**), taken from the CID, shows two female speakers (AB and CM) talking about their cinema-going habits and the different ticket types that are available. In this extract, CM (listener) performs an other-repetition of the words AB (storyteller) has just uttered (underlined). Taken together, the form of the repetition (the prosodic cues given by CM, the speech overlap and the changing of the deictic pronoun) and the context allow AB to identify the pragmatic function of this repetition as a confirmation request (Perrin *et al*., 2003 [11]). This is attested by the fact that AB treats CM's repetition as a confirmation request by recognising the requirement for a yes/no response and answering with the preferred "ouais" (circled) (Schegloff, 2007 [12]). This confirmation request repetition is used to display alignment, showing how both interlocutors are engaged in the same activity and cooperate in order to achieve a common goal[c]. In this case, the common goal is two-fold; the participants work together in order to adhere to the given task and, more generally, to maintain their relationship (here, as friends).

[c] For a more detailed analysis, see (Guardiola, 2014 [2]).

The following extract shows how the same phenomenon can be used in order to achieve a different common goal, in this case obtaining information. This extract shows an interaction between two tourists (T1 and T2) and a tourist advisor (C1) taken from MITo. C1 is answering questions from T1 and T2 concerning the various transport tickets in the city. Therefore, in this case, the common goal is external, in that the tourists are requesting information.

T2: et pour les tickets on le prend dans le bus	T2: and for the tickets we buy them on the bus
C1: avec le chauffeur directement	C1: with the driver directly
T1: [ah oui y a pas de c'est pas]	T1: [ah yes there isn't it isn't]
T2: [directement y a pas de lieu]	T2: [directly there isn't a place]
T1: c'est le bar tabac c'est pas [nécessaire]	T1: it's the tabacco kiosk it's not [necessary]
C1: [ah vous pouvez]	C1: [ah you can]
oui oui vous pouvez acheter au bar des tabacs dans les metros avec le chauffeur	yes yes you can buy them in the bar the kiosks in the metros with the driver
T1: avec le chauffeur	T1: with the driver
C1: oui	C1: yes

Figure 4. Example of confirmation request repetition in MITo in the original French (left) and a translation (right).

In much the same way as the previous example, T1 performs an other-repetition of part of C1's previous utterance (underlined). Once again, the form and context of the verbatim repetition allow C1 to identify this repetition as a confirmation request and respond accordingly (with the preferred "oui" (circled)), thereby displaying alignment[d] (cf. **Figure 4**).

As can be seen, certain elements are observable in both corpora such as the sequential organisation and pragmatic function of other-repetitions as well as their role in alignment and reaching a common goal. However, certain elements are only observable in one corpus or the other. In the CID, the prosodic cues alluded to in the example above are observable in detail thanks to the high-quality audio equipment used. The fact that each speaker is recorded on a separate channel allows this detailed analysis even in cases of speech overlap (such as the example above). Concerning MITo, similar analysis of prosodic cues is rendered almost impossible due to the presence of ambient background noise and speech overlaps on the same recording channel. Similarly, although not necessarily present in this particular example from the CID, the precise temporal alignment of speech (overlaps, delay, etc.) and the other modalities available (gesture, gaze, etc.) in the corpus, allows detailed analyses of embodied practices (Bertrand *et al*., 2008 [13]). This is not the case when working with MITo as the absence of video recording, due to restrictions on video equipment in the interactional setting, removes the possibility of analysing certain embodied practices. There are also elements that may be observed in MITo but not (or less so) in the CID. Firstly, MITo allows language use to be analysed in its "natural setting", something some researchers consider invaluable in analysing social interaction (Traverso, 2003 [8]). As shown in the example above, MITo permits analyses to be carried out on how verbal interactional resources (such as other-repetitions) are used to ensure achievement in specific tasks (such as requesting/providing information). Furthermore, in other examples, this natural setting allows the exploration and analysis of how participants use the "natural" communicational environment as a resource in interaction (for example, in this case, the use of leaflets, maps, etc.).

In conclusion, the two analyses presented here lead to a number of similar observations. To add to this, the analysis of the CID extract contributes elements which the MITo extract does not (and/or cannot) and vice versa. In this respect, these two analyses can be considered complementary. Furthermore, crossing these two analyses allows researchers to confirm the validity, in a natural setting (MITo), of observations made in a laboratory setting (the CID). Equally, observations made with the CID concerning precise analyses of multimodal and temporal information can potentially be used to shed light on similar phenomena observed in MITo, albeit with less "granularity". It is important to note here that these different yet complementary perspectives are made possible by the fundamental differences of the corpora, themselves indebted to the choice of different non-negotiables. For example, in the CID, the non-negotiable requiring access to as many sets of different

[d] For a more detailed analysis, see (Wilson, in preparation [1]).

resources as possible, in the finest detail possible, leads directly to the constitution of a corpus which then allows prosodic contours to be analysed even in cases of overlap. Similarly, MITo's non-negotiable of acquiring ecological data in a specific institutional context leads to a corpus which can be used to analyse how interlocutors use certain resources to accomplish specific tasks. This reveals a somewhat direct link - passing through *obligations*, *constraints* and *concessions* - between the choice of (different) non-negotiables and obtaining different perspectives which, in a context of good communication, leads to heightened possibilities of *a posteriori* interdisciplinary work.

6 Discussion

6.1 Audio recordings corpora considered as complex

As identified in section 4.2, the idea that corpus complexity should be defined solely on the number of modalities collected in the data may be brought into question. This definition implies that analytical access to a certain modality is dependent only on being able to record or capture said modality. For example, gesture may only be analysed if video recording is possible. However, it can be argued that certain aspects of multimodal complexity may be accessed through a single channel, especially that of audio recording. This argument is inspired by the pioneering work of Traverso (2012 [14]) on what she terms the "*univers sonore*" highlighting the possibility of accessing multimodal information through the analysis of other aspects of an audio recording away from simply the speech. Therefore, while care should be taken and, certain information concerning, for example, gesture, physical positioning of participants and use of objects or other semiotic resources can be gleaned from audible "traces" on a sound recording.

Though it should be noted that this possibility of accessing the "*univers sonore*" does not equate to a fine-grained multimodal analysis, in this light, "non-complex" audio recordings could indeed be considered "complex" as they provide a hive of information that goes beyond what is being said and how it is being said. As Traverso (2012 [14]) herself points out, work in this area is severely underdeveloped. This is perhaps due to huge improvements in technology allowing more and more modalities to be captured more and more easily. However, these same technological advances lead to ever-improving audio equipment that offers high quality recording capabilities in smaller and smaller packages which can be used in "ecological" settings. In the same way that this equipment allows more and more possibilities for (relatively) fine-tuned phonetic analysis of action or turn construction, the potential for enriching observations of audio recordings by taking into account the "*univers sonore*" is huge and would clearly benefit from further consideration.

6.2 To what extent can observations of a single corpus be generalised?

Another theoretical concern linked to the description of these corpora regards the generalisability of the phenomena observed. It can be argued that, with any corpus, the analysis of a given interaction leads to the identification of four types of phenomena: phenomena that can be considered transversal in all human interaction, phenomena linked to the general type of interaction being observed, phenomena originating from the specific characteristics of one particular interaction and, potentially, phenomena provoked by the presence of the linguist or recording equipment. The linguist must then judge to what extent each phenomenon observed can be generalised. This is true for any corpus and both corpora presented here display phenomena belonging to each category. For example, in both corpora participants apply rules concerning turn-taking, an observation which holds for any oral interaction (Sacks *et al*., 1974 [15]). Numerous observations of laughter in the CID may be linked to the type of interaction (oral face-to-face conversation) (Holt, 2010 [16]) in the same way that frequent other repair in MITo attests to an exolingual face-to-face interaction (Dausendschön-Gay, 1988 [17]).

More importantly, these considerations become especially delicate when working with controlled or semi-spontaneous corpora as the potential for phenomena to be influenced by the researcher and/or

the specific (and often unusual) characteristics of interaction is much greater. The interactions in MITo display many phenomena typically linked to service encounters, such as an abundance of polite formulae (as discussed by Kerbrat-Orecchioni (2006 [18]), among others). This can be attributed to the natural setting being a service encounter rather than the influence of the linguist's presence. In the CID, it could be argued that the high number of narratives in the corpus is linked to the conversational style of the interaction. However, it could also be argued that this phenomenon is due to the instructions given to the participants by the researcher. It is crucial to keep this information in mind when analysing and presenting such data.

Finally the influence of the researcher's or material's presence can often be identified through the thematisation of these elements by the participants (Traverso, 2015 [19]). This is relatively common in the CID yet remains extremely rare in the case of MITo. It must also be taken into account that a certain reticence or unwillingness on the part of a given participant to speak may be attributable to the recording situation itself. This phenomenon is much more easily observable in the CID than in MITo. Indeed, instances in the latter seem practically non-existent. However, further research taking into account the use of "proximity" or "distance" language (see Kock & Oesterreicher (1985 [20])) or different registers, genres and styles (see Biber & Conrad (2009 [21])) could be very valuable in further analysis of these corpora.

This final point, concerning the influence of the researcher and the research situation on the corpus constitution, leads to questions regarding the role of the researcher in the elaboration of her data. Space constraints prevent us from addressing this issue in great detail but we feel it is important to highlight briefly the importance of these questions. It is well established that social interaction is a constant process of co-construction on the part of the participants (Gumperz 1982 [22]). It is important to remember that the observed situation is a social interaction like (and unlike) any other and, therefore, the relationship between participant(s), recorder(s), material(s) and observer(s) is continually under construction (Cameron *et al*., 1993 [23], Mondada, 1998 [24]). This process is attested to by the examples of participant thematisation alluded to above. As this process is constantly evolving, it obviously has a strong influence on the data collected. It is therefore important to assess this potential impact during analysis and also to make this assessment explicit when concluding and presenting research. There is a growing body of work on the influence of the researcher and research situation, and the ways of taking this influence into account, (see Dupouy, this volume [25] for example) highlighting the importance of this issue.

The fact that these questions are primary concerns in any research design points to a link between the choice of corpus "non-negotiable(s)" and the position of the observer and her material(s) which is likely to have an effect on both the "complexity" and "spontaneity" of a given corpus. In a future research project, the present authors will explore the thematisations from both corpora in an attempt to improve understanding as to how the observer(s) and material(s) constitute veritable social actors in terms of interaction and how participants react to these actors, whether present, absent or imaginary.

7 Conclusion

This paper presented a practical and theoretical reflection on the constitution of two interactional corpora with a special focus on the issues of spontaneity and complexity. It has been shown that a key methodological consideration in spoken corpus construction is that of the playoff between spontaneity and complexity. Indeed, while not necessarily incompatible, it has been shown that balancing these two dimensions is often the driving force behind key decisions in research and corpus design. This paper proposed that these key decisions are hinged on non-negotiable elements which engender obligations, constraints and concessions in corpus constitution. Taking into account the researcher's individual influence on these non-negotiables, this explains how different corpora are used to explore similar general research questions.

The notion of spontaneity has also been questioned through showing that it can be considered as an association of a continuum of degree of control with the binary notion of linguist intervention. Similarly, complexity can be considered as resulting not only from increasing the number of

modalities collected in the data but also from other factors, such as the richness of interaction. Finally, three key theoretical concerns were addressed: the potential complexity of audio recordings, the generalisability of corpora and the co-construction of data.

It has been argued that more explicit communication of the non-negotiables in the constitution of corpora could promote a deeper understanding of corpora themselves as well as their analyses. Equally, it has been shown that this can lead to different yet complementary analyses focussed on similar research questions. This is crucial in the current climate of increased interdisciplinary work. Innovative interdisciplinary work is often achieved through bringing together diverse individuals or organisations to form interdisciplinary research teams. This "*a priori*" approach to interdisciplinarity is undoubtedly highly valuable and fruitful and should be supported. However, it seems that another approach, that of "*a posteriori*" interdisciplinarity work of the type described here, is often overlooked. This approach offers an alternative to combining potentially conflicting non-negotiables in research design by promoting increased communication, understanding and acknowledgement of differences in corpus design. This, in turn, leads to a deeper understanding of results how they are obtained. As with all interdisciplinary work, this helps all researchers on the road toward their shared goal of achieving better description of social interaction and human communication.

8 References

1. A. Wilson, *Interactions exolingues et sociolinguistique de la globalisation*, PhD thesis, Aix-Marseille Université (in preparation)
2. M. Guardiola, *Convergence en Conversation: la similarité linguistique comme indice d'alignement et d'affiliation,* PhD thesis, Aix-Marseille Université (2014)
3. A. Wilson, *Language in motion in the era of globalization*, presented at The Sociolinguistics of Globalization, University of Hong Kong, Hong Kong (2015)
4. F. Gadet, R. Ludwig, L. Mondada, S. Pfänder, A-C. Simon, Un grand corpus de français parlé: le CIEL-F. Choix épistémologiques et réalisations empiriques, Rev. Fr. Ling. Appl. **17(1),** 39-54 (2012)
5. L. Mondada, *Wo fahrte vous in die schweiz? Bricolages plurilingues en interaction à la frontière*, presented at Institut de plurilinguisme, Fribourg, Suisse (2013)
6. P. Mustaers & J.Swanenberg, Super-diversity at the margins? Youth language in North Brabant, The Netherlands, Sociolinguistic studies **6:1**, 65-89 (2012)
7. M. Tellier, Quelques orientations méthodologiques pour étudier la gestuelle dans des corpus spontanés et semi-contrôlés, Discours **15** (2014)
8. V. Traverso, *Analyse des interactions : questions sur la pratique*, HDR thesis, Université Lumière Lyon 2 (2003)
9. C. Bowern, *Linguistic Fieldwork. A Practical Guide*, Palgrave Macmillan, New York (2008)
10. F. Laurens, J.-M. Marandin, C. Patin & H. Yoo, The used and the possible: The use of elicited conversations in the study of Prosody. In Yoo, H-Y & Delais-Roussarie, E. (eds), Actes d'IDP 2009, ISSN 2114-7612 (2009)
11. Perrin, L., Deshaies, D. & Paradis, C. Pragmatic functions of local diaphonic repetitions in conversation, Journal of Pragmatics, **35**, 1843-1860. (2003)
12. Schegloff, E.A. The organization of preference/dispreference in Schegloff, E.A. *Sequence Organization in Interaction: A Primer in Conversation Analysis, Vol. 1*, Cambridge, Cambridge University Press 58-96 (2007)
13. R. Bertrand, P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde & S. Rauzy. Le CID-Corpus of Interactional Data-Annotation et exploitation multimodale de parole conversationnelle, TAL, **49**, 3, 105-134 (2008)

14. V. Traverso, Analyses interactionnelles: repères, questions saillantes et évolution, Langue Française, **3**/2012, **175**, p 3-17 (2012)
15. H. Sacks, E. Schegloff & G. Jefferson. A Simplest Systematics for the Organization of Turn-Taking for Conversation, Language **50**, 696-735 (1974)
16. E. Holt, The last laugh: hared laughter and topic termination, Journal of Pragmatics, **42(6),** 1513-1525. (2010)
17. U. Dausendschön-Gay, Particularités des réparations en situations de contact, *Echanges sur la conversation,* 269-283*,* C.N.R.S., Lyon (1988)
18. C. Kerbrat-Orecchioni, Politeness in small shops in France, Journal of Politeness Research, **2**, 79-103 (2006)
19. V. Traverso, *Entre conception, transcription et analyse: réflexions sur un "regard en alerte"*, presented at ICODOC, Lyon (2015)
20. P. Koch & W. Oesterreicher, Sprache der Nähe - Sprache der Distanz, Romanistisches Jahrbuch 36/85, 15-43 (1985)
21. D. Biber, S. Conrad, *Register, Genre, and Style*, Cambridge University Press, N-Y, (2009)
22. J. J. Gumperz, *Discourse Strategies*. Cambridge, Cambridge University Press (1982).
23. D. Cameron, E. Frazer, P. Harvey, B. Rampton & K. Richardson, Ethics, advocacy and empowerment: Issues of method in researching language, Lang Commun, **13-2**, 81-94 (1993)
24. L. Mondada, Technologies et interactions dans la fabrication du terrain du linguiste, Cahiers de l'ILSL, **10,** 39-68 (1998)
25. M. Dupouy, l'exercice de (re)présentation de soi lors d'un enquête de terrain: négociation? enjeu épistémologique? (this volume)