DOI: 10.20396/joss.v14i00.20389

COMPARISONS BETWEEN FACE-TO-FACE AND TELEPHONE SPEECH: RESEARCH POSSIBILITIES

ROCHA, João Victor Pessoa[1*]
VITAL, Átila Augusto Soares[2]

[1]Federal University of Minas Gerais – ORCID: https://orcid.org/0000-0002-6476-8671
[2] Federal University of Minas Gerais – ORCID: https://orcid.org/0000-0001-9875-4799

Abstract: *This pilot study aims at comparing two speech corpora in order to capture possible features of two different speech formats. The first corpus is a face-to-face speech corpus, the C-ORAL-BRASIL I (Raso and Mello, 2012) and the second corpus is a phone call corpus, part of the C-ORAL-BRASIL II (Raso; Mello; Ferrari, to appear), both in spontaneous informal Brazilian Portuguese. Through a pipeline prepared in Python, we compared: illocutionary complexity, informational density, overlapping, and disfluencies. Our results indicated that the speakers in the telephone calls had to use more vocal elements to convey a message while in face-to-face interactions the message would be divided in speech and gestures. The findings will be of interest to those who deal with speech corpora in different formats and are interested in selecting certain features for the comparison. Furthermore, this study has methodological ramifications because the elements examined here are very likely to occur in spontaneous speech in any format, and need to be compared carefully.*

Keywords: Speech corpora; Illocution; Telephone call; Disfluencies.

[*] Corresponding author: joaoprvictor@gmail.com

1 Introduction

This paper aims to explore research possibilities at the interface between pragmatics and informational structure using a corpus of Brazilian Portuguese telephone speech. The effects that different communication modalities have on the structuring of the code and the functions performed by linguistic units are well-documented in the literature. For instance, face-to-face oral communication is typically interactive and involves a range of multimodal resources, such as co-speech gestures, facial expressions, and the sharing of time and space between speakers. Written communication, on the other hand, is less interactive, as writers and readers do not share the same temporal or spatial context (Biber and Conrad, 2009).

Experiments have demonstrated that in communicative situations where speakers can see each other and use gestures, linguistic structures are adjusted to meet the listener's needs. Depending on the communication channel, speakers modulate the code and produce utterances with, for instance, reduced phonetic redundancy (Pate and Goldwater, 2015). Biber and Conrad (2009) distinguish a series of specific mediums of communication, such as telephone and radio in the oral channel, and handwritten letters and emails in the written register. Although language today manifests through various modalities, the oral channel – understood in its multimodal nature (including gestures and facial expressions) – remains the most prototypical form of human communication.

According to Moneglia (2011), channel variation has been analyzed in studies ranging from sociolinguistics to corpus annotation research. In this regard, the following parameters are relevant for analyzing the oral channel:

a. Face-to-face interactions in natural contexts
b. Telephone recordings
c. New media audiovisual interactions
d. Human/machine interactions
e. Media productions
f. Written-to-be-spoken communication

Our study seeks to compare dialogues produced under parameter (a) with those produced under parameter (b). Using two corpora of spontaneous speech – one capturing face-to-face interactions and the other telephone calls – annotated according to the Language into Act Theory (L-AcT) (Cresti, 2000; Moneglia and Raso, 2014; Cavalcante, 2020), we examine how illocutionary acts and different informational units in spoken discourse are distributed. Given that the corpora differ in terms of mediums of communication, it is expected that the informational structure and the number of speech acts (Austin, 1962) will also be modulated by the speakers to enhance communicative efficiency in each context (Levshina, 2022).

This paper is divided as follows: section 2 debates previous research done on telephone conversation language, while section 3 introduces a short description of each corpus. In section 4, the L-AcT is discussed in further details. Also, the methodological steps and decisions are explained in section 5. Finally, the results and conclusions are presented in section 6.

2 Speech corpus: telephone conversation

Speech corpus compilation has always been a challenge for corpus linguists. Some of the reasons for that are the ethical procedures, equipment, team training, and corpus architecture. Nonetheless, research groups put effort into making this type of corpus for various registers, dialects, languages, and settings. Being one of the speech interactions we do, telephone conversations are one of the possibilities for corpus source material that can offer insights especially regarding segmental and suprasegmental processes. Having said that, this section intends to describe some telephone conversation corpora and what they offer to the corpus linguistics community.

A corpus of Cantonese telephone interactions which amounts to approximately 200 hours of recording was compiled by Lo *et al.* (2001). The authors divided the corpus into two main parts: the first one is made of sentences, short paragraphs, and spontaneous conversation (with prompts); while the second part is related to pronunciation of named entities, foreign currencies, and navigation commands. The second subset is focused on developing the material for specialized technologies for certain areas. The corpus was able to cover 85% of all Cantonese tonal syllables. Considering the scarcity of resources for Cantonese, this corpus can be a valuable material for research and technology development.

Still within the scope of Asian languages, a large corpus of Mandarin telephone recordings was created by Liu *et al.* (2006). The corpus includes 2412 speakers, ranging from teenagers to elderly individuals, and covers a wide range of topics. Additionally, it contains 2,745,181 syllables and includes all 408 toneless base syllables. The team also made the metadata available, enabling sociolinguistic research, among other applications. An early evaluation showed that the data in the corpus improved accuracy in automatic speech recognition tasks, demonstrating its usefulness.

Telephone chats can also have a quite structured script, which was the case for the corpus of English telephone surveys presented in Camelin (2006). In this case, through a brief message, users were asked to contact a phone number to voice their satisfaction with the customer service they previously received. After the recording collection and evaluation, they proceeded to annotate the recordings in regards to the sentiment (neutral, positive, and negative), and to the topic of the opinion (courtesy, efficiency, rapidity, and other). They noticed that filled pauses, false starts, restarts, repetitions and discourse markers would impact the computational processing of the audio recordings. These phenomena also happen in face-to-face interactions, which can indicate that they are features of speech in any capacity and channel.

As for Portuguese, conversation openings like “alô” (hello) and greetings such as “oi” (hello) and “tudo bem?” (how are you?) are common in telephone and computer-mediated chats. In Almeida et al. (2014), the authors analyzed the prosody correspondence in a set of informal European Portuguese telephone recordings. The authors consider prosodic correspondence when the next speaker in the chat begins their turn with an f0 contour equal or similar to the previous speaker’s turn (Reed, 2012). They found out that prosody correspondence is, among other things, marking collaboration between speakers. This process corroborates what was discussed in Šturm (2021), which states that the closer the relationship between speakers, the sooner the prosodic correspondence may occur during a phone call.

Another use for telephone dialogue examination is in the forensic area. After analysing some samples of fake kidnapping and prize recharge scams done during phone calls, Pereira e Silva *et. al.* (2017) argue that tone, intensity, and f0 contour were the main phonetic components that mostly portray the image of a cruel, violent, and scary criminal. In a face-to-face kidnapping setting, there may be weapons and body language involved that would highlight the danger. However, in a phone call context, the (fake) kidnapper only has their voice to convey it.

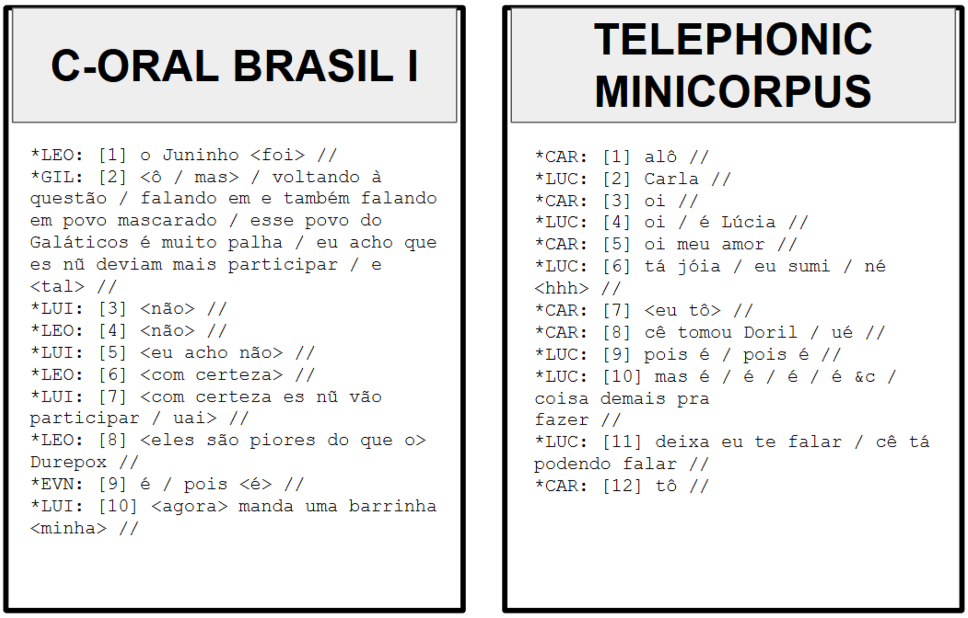
As demonstrated by the studies discussed in this section, phone calls can be quite revealing about the resources speakers use when they only have their voice to convey all meanings they desire to communicate. Because of that, our expectation was that in our sample we would find elements that would confirm this, especially when compared to in-person interactions.

In the next section you will read a description of the corpora we analysed.

3 C-ORAL-BRASIL I and the Telephonic minicorpus

In order to carry out the analysis, the following corpora were used: C-ORAL-BRASIL I (Raso and Mello, 2012) and the Telephonic minicorpus, part of the C-ORAL-BRASIL II (Raso; Mello; Ferrari, to appear), both in spontaneous informal Brazilian Portuguese (pt-BR) from Belo Horizonte, Minas Gerais, Brazil. The C-ORAL-BRASIL I and the Telephonic minicorpus are available online[1] alongside other corpora that are products of the C-ORAL-BRASIL project. The C-ORAL-BRASIL II corpus is undergoing final preparations before publication and comprises formal, media, and telephone interactions. Moreover, both corpora provide prosodic segmentation, high acoustic quality, and annotation of lexicalization and grammaticalization processes in pt-BR.

C-ORAL-BRASIL I (Raso and Mello, 2012) features 139 texts divided in different settings (private and public) and typologies (monologue [one speaker], dialogue [two speakers], and conversation [three or more speakers]). Furthermore, this corpus has a high diaphasic variation, i.e. variation in the communicative situation, at the same time maintains a balance in sociolinguistic parameters, more specifically sex, schooling level and age. On the other hand, the Telephonic minicorpus is a branch in the C-ORAL-BRASIL project that documents telephone calls. All 27 texts in the corpus also cover an array of various situations, for example, a customer requesting home services and a granddaughter describing the new house to her grandmother. Figure 1 displays extracts from both corpora and their audio files are also available.



C-ORAL BRASIL I

```
*LEO: [1] o Juninho <foi> //
*GIL: [2] <ô / mas> / voltando à
questão / falando em e também falando
em povo mascarado / esse povo do
Galáticos é muito palha / eu acho que
es nũ deviam mais participar / e
<tal> //
*LUI: [3] <não> //
*LEO: [4] <não> //
*LUI: [5] <eu acho não> //
*LEO: [6] <com certeza> //
*LUI: [7] <com certeza es nũ vão
participar / uai> //
*LEO: [8] <eles são piores do que o>
Durepox //
*EVN: [9] é / pois <é> //
*LUI: [10] <agora> manda uma barrinha
<minha> //
```

TELEPHONIC MINICORPUS

```
*CAR: [1] alô //
*LUC: [2] Carla //
*CAR: [3] oi //
*LUC: [4] oi / é Lúcia //
*CAR: [5] oi meu amor //
*LUC: [6] tá jóia / eu sumi / né
<hhh> //
*CAR: [7] <eu tô> //
*CAR: [8] cê tomou Doril / ué //
*LUC: [9] pois é / pois é //
*LUC: [10] mas é / é / é / é &c /
coisa demais pra
fazer //
*LUC: [11] deixa eu te falar / cê tá
podendo falar //
*CAR: [12] tô //
```

Figure 1: Extracts from the corpora (audio files available at: https://bit.ly/41WrflS)

4 Language into Act Theory, information units and the telephonic node of C-ORAL-ROM

Representative samples from the corpora of the C-ORAL family have been informationally annotated following the Language into Act Theory (L-AcT) (Cresti, 2000; Moneglia and Raso,

[1] https://www.c-oral-brasil.org/corpora_para_download.php

2014; Cavalcante, 2020). L-AcT is a corpus-driven theory, formulated based on decades of analysis of spontaneous speech corpora. It represents a contemporary development of Speech Act Theory (Austin, 1962), integrating prosodic analysis and informational structure into the study of communicative sequences. Its basic units of analysis are utterances and so-called stanzas (check examples 1 and 2 below). The former refers to prosodically complete linguistic sequences that convey a single illocution, while stanzas are also prosodically complete sequences but convey two or more illocutions.

The sound signal of prosodically complete linguistic sequences can be segmented into prosodic units with non-terminal breaks. These breaks delimit tonal units, which in turn correspond to different informational units (Cresti and Moneglia, 2010). Illocutions in utterances are conveyed through Comment units (COM), whereas in stanzas, they can occur as Bound Comments (COB) and Multiple Comments (CMM) in addition to the COM unit itself.

In the course of speech, speakers may choose to produce other informational units that do not convey illocutions but provide information and cognitive elements to aid the interpretation of illocutionary units. The non-illocutionary units that add locutive content to an utterance are the following: Topic unit (TOP), Appendix of Comment (APC), Appendix of Topic (APT), Parenthetical (PAR), and Locutive Introducer (INT). There are also dialogic units, which do not add information to the text of the linguistic sequence but serve to regulate the interaction: Incipit (INP), Expressive (EXP), Allocutive (ALL), Conative (CNT), Discourse Connector (DCT), and Evidentiator (EVD) (Cavalcante, 2020).

Example 1 represents an utterance composed of two tonal units. The first unit realizes a Topic informational unit (TOP), whose function is to establish the cognitive domain for the application of the illocutionary force that follows. The illocution in the next unit is tagged with the Comment (COM) and exhibits a terminal prosodic profile (figure 2).

Example 1 (bpubdl01)
*BRU: as recarregáveis /=TOP= tão aqui //=COM=
BRU: the rechargeable ones /=TOP= are here //=COM=

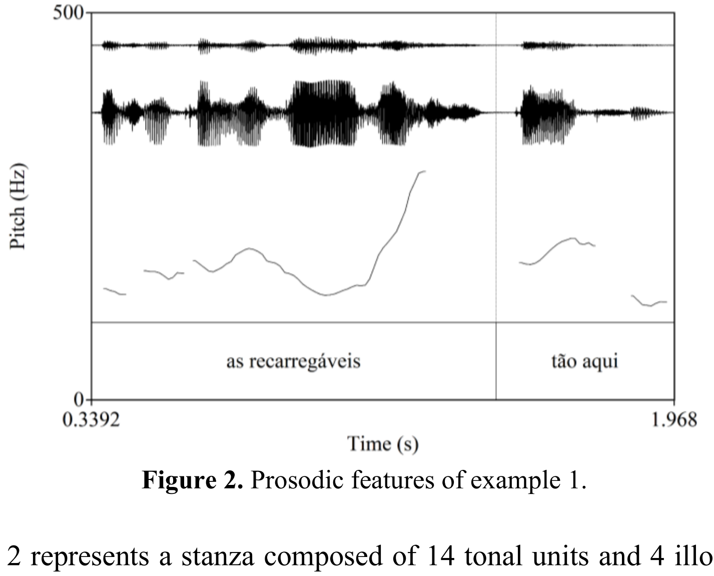


Figure 2. Prosodic features of example 1.

Example 2 represents a stanza composed of 14 tonal units and 4 illocutions. The TMT units refer to filled pauses. In the annotation system used, disfluency phenomena are also represented, such as word retractings ([/]), interruptions (+), and fragmented words (&).

Example 2 (bpubdl01)

BAL: existem vários /=COB= só que a maioria /=TOP= &he /=TMT= tá julgando improcedência /=COB= tal /=COB= porque /=DCT= &he /=TMT= de certa forma /=TOP= a bancada evangélica /=TOP= eles tão /=SCA[2]= muito contra /=COM= essa coisa /=APC= né //=EVD=

BAL: there are several /=COB= but most /=TOP= &he /=TMT= is ruling them unfounded /=COB= like /=COB= because /=DCT= &he /=TMT= in a way /=TOP= the evangelical caucus /=TOP= they are /=SCA= very against /=COM= this thing /=APC= you know //=EVD=

The greater the number of illocutions in a stanza, the higher its potential for complexity tends to be (Rocha et al., 2022). This is because informational subpatterns can form between one illocutionary unit and another, providing cognitive and semantic-pragmatic domains for interpreting the illocution governing that pattern. The methodological steps for understanding the informational density and illocutionary complexity of the telephone corpus are based on these internal pattern factors, as will be presented in the next section.

Within the scope of L-AcT, lexical, morphosyntactic, and semantic-pragmatic properties are defined concerning reference units (the utterance and the stanza). Cresti (2005) presents a series of lexico-structural strategies defined based on the utterance for the description of the C-ORAL-ROM corpus (Cresti and Moneglia, 2005), a spontaneous speech corpus of the four main Romance languages (Spanish, Portuguese, Italian, and French). The description also considers the telephone part of the corpus, which, like face-to-face interactions, is prosodically segmented.

In the L-AcT framework, so-called simple utterances are those containing a single prosodic unit that conveys a single Comment informational unit. Compound utterances, on the other hand, have at least one non-terminal break and, in addition to the Comment unit, at least one other informational unit. In Cresti (2005), lexico-grammatical strategies are presented in light of these two types of terminated sequences. One such strategy is the distinction between utterances (simple or compound) that include finite verbs (verbal utterances) and those that do not (verbless utterances). The presence of a verbal phrase around which much of the linguistic structure is organized within an utterance makes this type of sequence more complex than those not based on finite verbs. Examples 3–6 from Cresti (2005) below illustrate these observations.

Example 3 (ifamcv10)
Compound verbless
*LUC: il gelato / no //
**LUC: ice cream / no way //*

Example 4 (ifamcv10)
Simple verbal
*LUC: oggi fa freddo //
**LUC: it's cold today //*

Example 5 (ifamdl08)
Simple verbless
*ELA: tutto il giorno //
**ELA: all day long //*

[2] SCA stands for scanning.

Example 6 (ifamdl16)
Compound verbal
*CLA: quando lei va via la sera / nell'ascensore 'un ce più luce //
CLA: when she goes away in the night / in the elevator the light is off //

Based on observations from the telephone section of the Italian language in the C-ORAL-ROM corpus, Cresti (2005) highlights several relevant points. The telephone corpus contains the highest number of simple utterances without a verbal function (as in Example 5 above). The same trend can be observed in the telephone sections of the European Portuguese and Spanish languages. On the other hand, the telephone section is the one that, among all texts and languages in the corpus, presents the lowest number of compound utterances with a verbal function (Example 6).

Although not the focus of this study, it is important to keep in mind that, within the L-AcT framework itself, telephone interactions have particular characteristics compared to other types of interactions available in spontaneous speech corpora. According to Cresti (2005, p. 214), telephone situations "represent the lowest step of speech structuring".

Building on the tradition of preparing telephone corpora within the C-ORAL family, the proposed methodology will address the complexity of sequences and the nature of illocutionary units. Naturally, this is only one of many possible approaches to analyse the Telephonic minicorpus, considering the other resources available for the description of Portuguese developed by the Laboratory for Empirical and Experimental Studies on Language (LEEL) at the Faculty of Letters, UFMG (Brazil).

5 Methods

This section presents the methodological steps employed in this study. This is a preliminary investigation aiming to explore the potential uses of two corpora compiled within the scope of the C-ORAL-BRASIL project. We conducted a comparison between corpora documenting face-to-face and telephone interactions. The face-to-face corpus analysed here contains 46 texts extracted from the dialogic typology section of C-ORAL-BRASIL I, while all texts from the Telephone minicorpus are examined here.

The choice of the dialogic section of C-ORAL-BRASIL I for comparison with the telephone corpus is justified by its higher comparability to telephone interactions. In most cases, during a phone call, dialogue (i.e., two speakers talking with each other) is the predominant textual typology that emerges. In a comparison, potential differences in linguistic patterns between telephone and face-to-face interactions can be attributed to (i) the structure of the speech event (monologue, dialogue, or conversation), (ii) the nature of the communication channel, (iii) the sociological domain (family/private or public), and (iv) the formal or informal register. Other elements, such as text genre, sociolinguistic factors, communicative purpose, and topic naturally also influence linguistic characteristics (Mello, 2014). However, studying these elements requires specific methodologies distinct from the one proposed here.

Spontaneous speech, i.e., speech planned while it is produced, predominantly occurs through multimodal face-to-face interactions. The shared space and time in such cases enable the use of a wide variety of gestures, anticipating that the audience will perceive them and interpret the message in the most efficient way (Levshina, 2022). Spontaneous speech can also occur via electronic communication channels, such as telephones and cell phones (as in the case of the telephone corpus) and, more recently, through online meeting and video call technologies. In telephone calls, speakers do not share the same spatial references or maintain visual contact. Through the telephone channel, only the content conveyed via sound waves must package the necessary information and facilitate communicative exchanges (Mello, 2014).

This exploratory study of the differences observed in telephone chats compared to face-to-face interactions considered measures of illocutionary complexity, informational density and speech overlapping. To present the methods used to extract these measures, we will consider the theoretical framework guiding the informational annotations in both C-ORAL-BRASIL I and the telephone corpus: the Language into Act Theory (Cresti, 2000; Moneglia and Raso, 2014; Cavalcante, 2020), already introduced in section 4.

5.1. Illocutionary Complexity and Informational Density

A series of methodologies can be applied in the comparison of the telephone corpus with a face-to-face interaction corpus, but each has its particularities defined by the research objectives. A comparison of lexical items, for example, is highly sensitive to the topics discussed, while an analysis based on the classification of illocutionary types reflects the action-oriented goal of the interactions (giving instructions, making requests, providing information, etc.). For instance, a comparison based on informational subpatterns allows us to examine how telephone interactions convey more or fewer pieces of information for each illocution compared to face-to-face interactions. By utilizing the informational annotations from both corpora, our methodology enables the comparison of the number of illocutions at different complexity levels and the amount of information conveyed in the linguistic sequences.

To make these measurements, we base our approach on the methodological framework of Rocha et al. (2022) for studying informational complexity in the speech of patients with schizophrenia through comparisons between C-ORAL-ESQ (Raso et al., 2024), a corpus that documents the speech of patients with schizophrenia, and C-ORAL-BRASIL I. In the absence of a control corpus – with the same communicative situations as the psychiatric consultations documented in C-ORAL-ESQ – a methodology was presented that compares terminated sequences with the same number of illocutionary units between the two corpora. The validity of the proposal is based on the idea that sequences with the same number of illocutions have comparable potentials for informational complexity (Rocha et al., 2022).

We refer to illocutionary complexity as the number of terminated sequences (utterances and stanzas) with different amounts of illocutionary units. This number was calculated based on a random sampling of 100 completed sequences from each corpus. With this measure, it is possible to verify which channel facilitates the exchange of the greatest number of illocutions. Similarly, we define informational density as the ratio between the number of non-illocutionary informational units and the number of illocutionary units in a given linguistic sequence. The higher the informational density, the greater the number of non-illocutionary units that provide the conditions for interpreting the illocution in the pattern. This measure was also conducted for 100 randomly collected utterances from both corpora.

All methodological steps were performed automatically using Python scripts[3]. A data frame was programmed to obtain the numbers of illocutionary and non-illocutionary units for each linguistic sequence. In a subsequent step, interrupted utterances, EMP (empty) units generated by disfluencies, and SCA (Scanning) units generated by the lack of isomorphism between informational and tonal units were removed. The flowchart in Figure 3 presents the steps carried out up to the separation of sequences with different numbers of illocutionary units. The

[3] Available at https://github.com/joaoprvictor/telephone_oral_analysis

random sampling collects sequences from each corpus and checks their illocutionary complexities and informational densities.

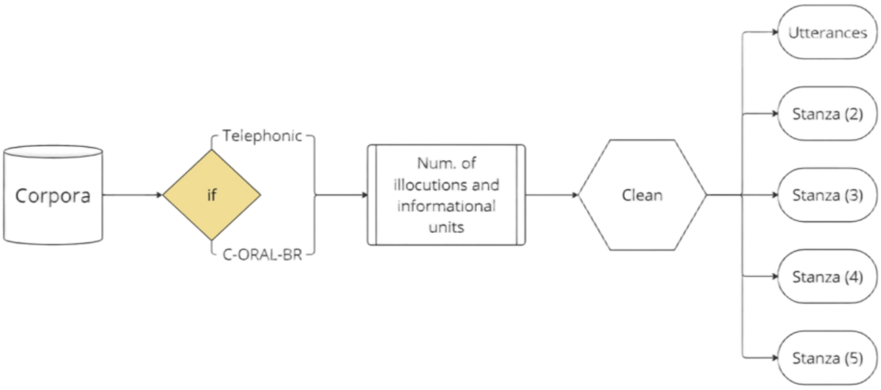


Figure 3. Flowchart of the procedures.

As an example, let us calculate these metrics for Examples 1 and 2 reported above. Table 1 provides information on the number of illocutionary units and informational units for the two linguistic sequences. The informational density is calculated from the ratio between the fourth and third columns.

Table 1. Example of the measurements for the sequences of examples 1 and 2.

Dataframe				
Sequence	Taggs	Num. of ill. units	Num. of info. units	Informational density
Ex. 1	[TOP, COM]	1	2	2
Ex. 2	[COB, TOP, TMT, TMT, COB, COB, DCT, TMT, TOP, TOP, SCA, COM]	4	11	2.75

It can be observed that example 1 has an illocutionary complexity of 1 and an informational density of 2, as it contains one illocution for two informational units; example 2 has an illocutionary complexity of 4 and an informational density of 2.75, as it includes four illocutions for eleven informational units[4]. With the same data frame, measures were extracted on the rate of overlapping between the speakers. The methods used are presented in the next subsection.

5.2. Overlapping and Disfluencies as a Metric of Interactivity

In addition to the two metrics for evaluating the structure of information, measurements were also made for the analysis of speech overlapping. Overlap refers to cases where two or more speakers produce linguistic material at the same time (Schegloff, 2000). The phenomenon of overlapping is another typical characteristic of sharing the same speech production time among speakers. In written registers, for example, this phenomenon does not occur, as the time of the author is different from the time of the recipient, and there is no competition for the conversational turn (Vinciarelli; Chatziioannou; Esposito, 2015).

[4] The SCA is not considered an informational unit (Cresti, 2000).

Both in the telephone corpus and in C-ORAL-BRASIL I, overlap is marked with the symbols “<” and “>”, which represent the beginning and the end of the overlapping speech, respectively. For comparison, we calculated in the same data frame the number of overlapping words for each linguistic sequence. The result is given by the number of overlapping words in relation to the total number of words in the sequence (for example, a 10-word utterance with 4 overlapped words has 40% of its linguistic material overlapped).

The frequencies of four types of disfluencies were also described for both corpora: retractings ([/]), interruptions (+), interrupted words (&), and filled pauses (&he). Interruptions and interrupted words occur when the speaker abandons an entire linguistic sequence or the pronunciation of a word, respectively. Retractings, on the other hand, involve reformulations without abandoning the entire sequence (see example 7). They are usually motivated by mispronunciations, word repetitions, lexical selection errors, and other related phenomena (Vital and Rocha, to appear).

Example 7 (bfamcv04)
*BRU: cê não pode aprontar [/1] apontar pra mesa //
**BRU: you can’t proint [/1] point to the table //*

In Example 7, the speaker produced an utterance with a retracting triggered by the mispronunciation of a word ('aprontar' instead of 'apontar'). In this case, rather than abandoning the entire utterance, the speaker produces a prosodic break and reformulates only the mispronounced word.

6 Results and conclusion

This section presents the results of the analyses of illocutionary complexity, informational density, overlaps, and disfluencies. As this is a preliminary study, the discussions surrounding the results aim to illustrate the potential applications of the resources provided by the telephonic corpus.

6.1. Illocutionary Complexity and Informational Density

The comparison between the corpora regarding illocutionary complexity was conducted using 100 samples of terminated sequences collected from the dialogue section of the C-ORAL-BRASIL I corpus and 100 samples collected from the texts of the Telephonic minicorpus. Figure 4 shows that face-to-face interactions and telephone interactions tend to favor sequences with lower complexity potential (fewer illocutionary units). In the case of the C-ORAL-BRASIL corpus, more than 70% of the samples consisted of a single illocution. Utterances of this complexity represented approximately 67% of the samples in the telephone corpus.

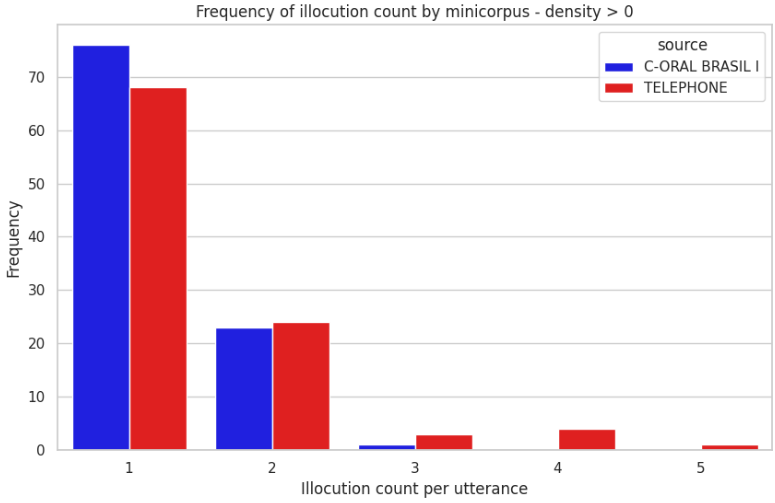


Figure 4. Frequency of terminated sequences with 1-5 illocutionary units.

It is important to note that no sequences with more than three illocutionary units were captured in the sample dialogues of the C-ORAL-BRASIL corpus. In terms of complexity, this result suggests that, in face-to-face dialogue interaction, there appears to be a tendency for speakers to produce sequences with lower potential for complexity. Unlike the measures presented by Cresti (2005), the potential for complexity in this case does not arise from the fact that the utterance is compound or simple and contains a significant verb or not. Here, the complexity depends on the potential number of subpatterns that can be formed with a given number of performed illocutions.

A similar complexity pattern has been observed in comparisons of different typologies within the C-ORAL-ROM (Cresti and Moneglia, 2005): the greater the interactivity required by the structure of the speech event (monologue and dialogue, with the latter being more interactive), the fewer the complex utterances tend to be (Cresti, 2005). This may occur because, in dialogues, the demand for turn-taking is higher than in monologues. As a result, speakers have less time to develop their conversational contributions without being interrupted or having their speech overlapped.

Example 8 was randomly selected to compose the sample of C-ORAL-BRASIL I. It is an utterance composed of three informational units, two of which are completely overlapped. The speaker and listener sharing the same physical environment enables the indexical “aqui” (here) to be used initially to refer to one location and then immediately to another. Unfortunately, there is no video recording of this interaction, but it is reasonable to assume that CES also performs co-speech gestures, i.e., gestures that co-occur with speech, of a deictic nature. The scope of the spatial reference can only be efficiently recovered by the interlocutor who visually follows the speaker’s gestures.

Example 8 (bfamdl05)
*CES: <aqui> /=TOP= ele é maior /=COM= <do que aqui> //=APC=
*CES: <here> /=TOP= it is greater /=COM= <than here> //=APC=

In Figure 4, utterances consisting of a single Comment unit were excluded, i.e., utterances with an informational density equal to zero (0). The same measure, including these utterances, was applied to another sample of 100 terminated sequences per corpus (Figure 5).

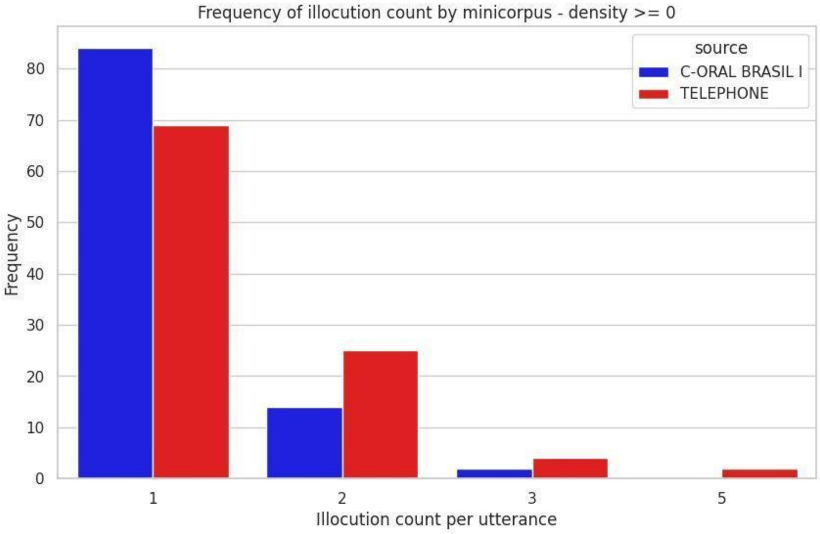


Figure 5. Frequency of sequences with 1-5 illocutionary units, considering utterances with informational density >= 0.

In this sample, the pattern of a higher number of utterances with a single illocutionary unit from the C-ORAL-BRASIL I corpus was confirmed compared to Figure 4. The same holds for stanzas with two, three, and five illocutionary units, for which there are more samples from the telephone corpus, reinforcing the observation that there are more illocutionary complex sequences in this corpus (Figure 5). In this second sample, no data with four illocutionary units were found.

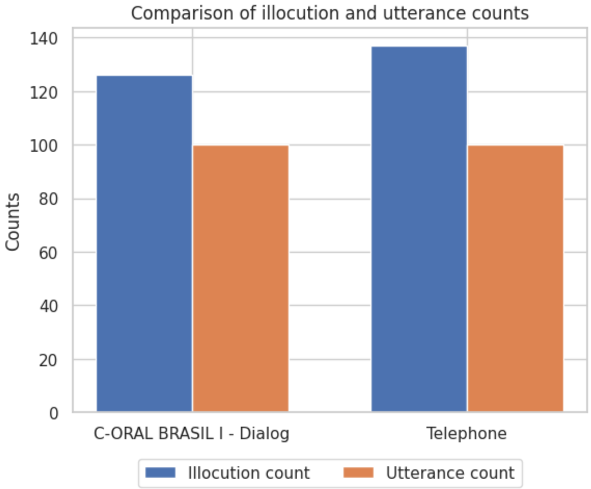


Figure 6. Number of illocutionary units found in the samples for each corpus.

Figure 6 shows that the total number of illocutions for samples of 100 terminated sequences also differs between the corpora. While the face-to-face interaction set has just over 120 illocutions per 100 utterances, the telephone set has nearly 140. This measure should be analysed in constant comparison with the informational density data, which are presented next.

One interesting feature of telephone chat is that almost all of them start with conversation openings, such as hello, what's up, who is this. These openings tend to occur as conventionalized illocutions and they provoke specific reactions. For example: if caller A says "hello" (a greeting illocution), it is very likely that caller B will respond with "hi[5]" (another greeting illocution). This unique characteristic may put the total count of illocutions in the corpus with a larger number. Examples of this type of phenomenon in the telephone corpus can be observed below, in examples 9 and 10, featuring sets of standardized illocutions with Multiple Comments (CMMs). Example 9 captures the first few seconds of a call, during which the speakers greet and identify themselves, whereas example 10 depicts the final seconds of a call, where the speakers exchange thanks and bid farewell.

Example 9 (btelpv02)

*KEN: alô //=COM=
***IMA: alô /=CMM= Kênia //=CMM=**
*KEN: oi //=COM=
*KEN: é Imaculada //=COM=
*IMA: é //=COM=
*IMA: tudo bem //=COM=
***KEN: tudo bem /=CMM= e você //=CMM=**

**KEN: hello //=COM=*
****IMA: hello /=CMM= Kênia //=CMM=***
**KEN: hi //=COM=*
**KEN: it is Imaculada //=COM=*
**IMA: yes //=COM=*
**IMA: all good //=COM=*
****KEN: all good /=CMM= and you //=CMM=***

Example 10 (btelpb01)

***CAR: tá marcado /=CMM= <viu> //=CMM=**
*LUC: <tá> oquei //=COM=
***LUC: <brigada> /=CMM= tá Carla //=CMM=**
*CAR: <beijo> //=COM=
*CAR: brigada ocê //=COM=
*CAR: <tchau> //=COM=
***LUC: <tchau> /=CMM= um abraço //=CMM=**
*CAR: outro //=COM=
*LUC: tchau //=COM=

****CAR: it's scheduled /=CMM= <got it> //=CMM=***
**LUC: <it's> okay //=COM=*
****LUC: <thanks> /=CMM= okay Carla //=CMM=***
**CAR: <kiss> //=COM=*
**CAR: thanks to you //=COM=*
**CAR: <bye> //=COM=*
****LUC: <bye> /=CMM= a hug //=CMM=***
**CAR: same to you //=COM=*
**LUC: bye //=COM=*

[5] Or any other conversation opening marker.

Both the greetings in example 9 and the farewell and gratitude ritual in example 10 seem to favor sets of illocutions with standardized prosodic forms (CMMs). Unlike illocutions performed with Bound Comments (COBs), those realized with Multiple Comments (CMMs) are conventionalized and represent illocutionary acts that are holistically understood by interlocutors. In contrast to sequences performed with COB-COM, the CMM-CMM pattern should not be interpreted as a composition of distinct utterances, but rather as a sequence of illocutions whose prosodic pattern is conventionalized to convey specific actions, such as lists, comparisons, reinforcements, and alternative questions (Moneglia and Raso, 2014).

It appears that the highly ritualized nature of conversational openings and closings conditions the realization of similarly patterned illocutions. A detailed corpus-based study on this subject remains to be conducted, and the Telephonic corpus provides all the necessary resources for such an endeavor.

6.2. Disfluencies and Overlapping

The disfluencies counted were filled pauses, interrupted words, sequence interruptions, and word retractings, with their respective frequencies shown in Figure 7. These metrics were calculated considering all the words in the corpora and may be useful for future characterizations of the telephone corpus. It is expected that, due to the longer duration of dialogical interactions (in both time and word count), the total number of disfluencies would also be higher for this typology and channel. However, the phenomenon of retractings is what results in the C-ORAL-BRASIL I corpus having significantly more disfluencies than the telephone corpus.

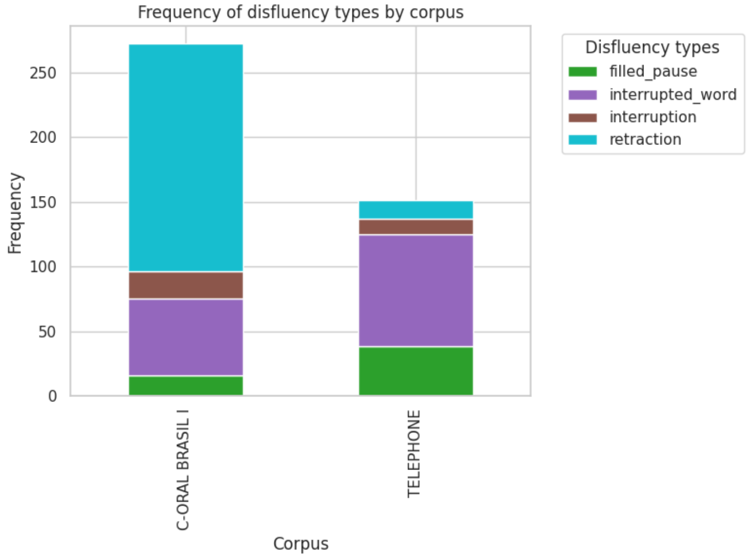


Figure 7. Quantification of disfluencies in both corpora.

Regardless of data normalization or sampling, it is also important to note the high frequency of filled pauses and interrupted words in the telephone chat corpus. The number of interrupted utterances, on the other hand, is not higher than that in the dialogues of the C-ORAL-BRASIL I corpus, but it is still quite significant, given the smaller number of words in the telephone corpus.

At this stage, the measure of disfluencies should be interpreted with caution, as new corpus-driven classifications have pointed to different types of phenomena that are often annotated under the same tag (Kosmala, 2024; Vital and Rocha, in press). This is particularly true for cases annotated as retractings, which may group repetitions, repairs, and changes in planning – phenomena of different natures – under the same retracting annotation.

The counting of overlapping was calculated for the samples from both corpora (100 terminated sequences each) and normalized according to the number of words present in the utterances. This means that the number of overlapped words represents a percentage of the total number of words in that utterance or stanza produced by the speaker. The boxplots in Figure 8 represent this result, considering only sequences with some part overlapped.

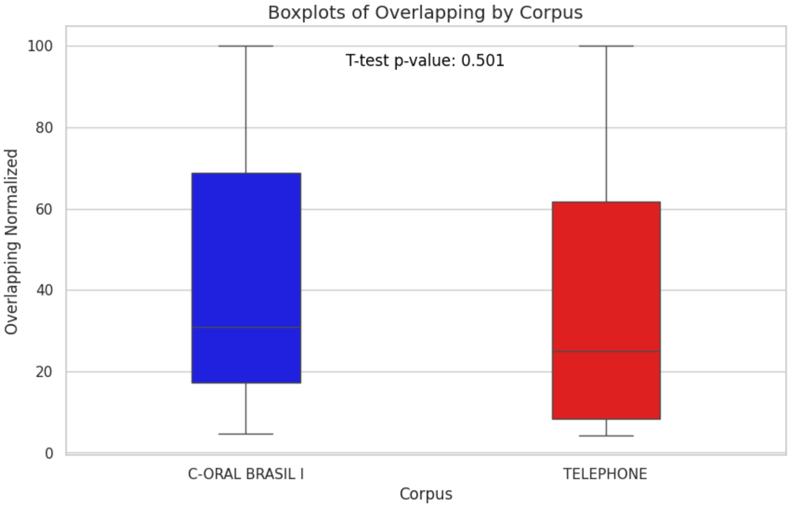


Figure 8. Boxplots of the percentage of overlapped words by corpus.

From both corpus samples, 29% of the sequences had overlapping (represented in figure 8). It should be emphasized the difference between the distributions of overlap percentages in the telephone corpus and in C-ORAL-BRASIL I. A relatively smaller number of words in the telephone channel is overlapped compared to face-to-face dialogues. The Student's t-test does not indicate statistically significant differences in these distributions (p = 0.501).

This project was undertaken to design possibilities of comparison between telephone and in-person conversation corpora and to evaluate how different they may be. Firstly, considering all the annotation markings that there are in the transcriptions, it is possible to contrast such corpora because they underwent the same process of transcription, prosodic segmentation and annotation, which ensures comparability. Furthermore, this study has found that generally speakers in face-to-face interactions tend to say utterances with lower complexity considering the informational density, whereas callers in telephone chats may say utterances with higher informational complexity.

The fact that the telephone chats samples presented more illocutions than the in-person interactions may have emerged as a predictor of how much a speaker must act in order to convey the message. In face-to-face dialogues, the speakers can use facial expressions, gestures, and geographical location in order to communicate. However, in a phone call, the speaker only has their voice to do so, which may implicate in having more illocutions as a way of balancing out with the other resources not used (body language being the main one).

As previously discussed, the results regarding disfluencies are up to debate since there has been some revisions of what a disfluency is and what they mark. Nevertheless, considering the telephone corpus, the interrupted words were a major disfluency. This may be due to the fact of technical issues, the phone network connection may be lagging, or of interactional processes, the caller may have more chances of correcting themselves. This result should be more deeply checked in further studies.

Overall, this study strengthens the idea that speech in whatever format will maintain certain features while other elements will be highlighted, for instance, depending on the channel of communication. Moreover, this study has methodological implications considering that the aspects analysed here are highly prone to happen in spontaneous speech in any form and need to be compared carefully.

The main weakness of our paper is the lack of robustness in the results due to the scarcity of more data, especially with the telephone conversation corpus. Despite its exploratory nature, this pilot study offers some insights that can instigate more research concerning comparisons of different speech corpora. For that reason, future studies should consider analysing more data and other languages as well as putting effort into annotating other linguistic interactional phenomena, such as turn-taking, Multiple Comments and discourse flow.

Acknowledgements

Since this is a journal volume dedicated to the amazing work done by prof. Heliana Mello, we could not miss the opportunity to express our gratitude to what she has contributed to science. Besides being super insightful and resourceful, Prof. Heliana Mello is an example of work ethics, intelligence, kindness, and empathy. Moreover, she has impacted the life of so many students, including the authors of this paper, by always encouraging us to be at our best. She has built a legacy of splendid science and warm rapport building and it will last for a long time even after her retirement.

REFERENCES

1. Almeida ANS de, Musiliyu O, Santana de Almeida RA, Oliveira JR. M. Correspondência e não correspondência prosódicas em aberturas de conversas telefônicas no português europeu. Rev Leitura [Internet]. 2014 Oct 11;2(52):293–316. Available from: https://www.seer.ufal.br/index.php/revistaleitura/article/view/1484

2. Austin JL. How to do things with words. Oxford: Oxford University Press; 1962.

3. Biber D, Conrad S. Register, genre, and style. Cambridge: Cambridge University Press; 2009.

4. Camelin N, Damnati G, Béchet F, De Mori R. Opinion mining in a telephone survey corpus. In: Interspeech; 2006 Sep.

5. Cavalcante FA. The information unit of Topic: a crosslinguistic, statistical study based on spontaneous speech corpora [PhD thesis]. Belo Horizonte: Universidade Federal de Minas Gerais; 2020.

6. Cresti E. Notes on lexical strategy, structural strategies and surface clause indexes in the C-ORAL-ROM spoken corpora. In: Cresti E, Moneglia M, editors. C-ORAL-ROM: Integrated reference corpora for spoken romance languages. Amsterdam/Philadelphia: John Benjamins; 2005.

7. Cresti E, Moneglia M. C-ORAL-ROM: Integrated reference corpora for spoken romance languages. Amsterdam: John Benjamins Publishing; 2005.

8. Cresti E, Moneglia M. Informational patterning theory and the corpus-based description of spoken language. The compositionality issue in the topic-comment pattern. In: Bootstrapping information from corpora in a cross-linguistic perspective. Firenze: Firenze University Press; 2010. p. 13–46.

9. Cresti E, et al. Corpus di italiano parlato. Introduzione e Campioni (II Voll.). Firenze: Accademia della Crusca; 2000.

10. Kosmala L. Beyond Disfluency: The interplay of speech, gesture, and interaction. Amsterdam: John Benjamins; 2024.

11. Levshina N. Communicative efficiency. Cambridge: Cambridge University Press; 2022.

12. Liu Y, Fung P, Yang Y, Cieri C, Huang S, Graff D. HKUST/MTS: A very large scale Mandarin telephone speech corpus. In: Chinese Spoken Language Processing: 5th International Symposium, ISCSLP 2006, Singapore, December 13–16, 2006. Proceedings. Berlin Heidelberg: Springer; 2006. p. 724–35.

13. Lo WK, Ching PC, Lee T, Meng H. Design, compilation and processing of CUCall: a set of Cantonese spoken language corpora collected over telephone networks. In: Proceedings of Research on Computational Linguistics Conference XIV; 2001 Aug. p. 193–212.

14. Mello H. Methodological issues for spontaneous speech corpora compilation. The case of C-ORAL-BRASIL. In: Spoken Corpora and Linguistic Studies; 2014. p. 27–68.

15. Moneglia M. Spoken corpora and pragmatics. Rev Bras Linguist Apl. 2011;11:479–519.

16. Moneglia M, Raso T. Notes on the language into act theory. In: Spoken corpora and linguistic studies. Amsterdam: John Benjamins Publishing Company; 2014. p. 468–95.

17. Pate JK, Goldwater S. Talkers account for listener and channel characteristics to communicate efficiently. J Mem Lang. 2015;78:1–17.

18. Pereira e Silva W, de Souza Melo MS, Gomes do Vale RP. Prosódia e construção de ethé discursivos em crimes via telefone. Fórum Lingüístico. 2017 Oct 1;14(4).

19. Raso T, Mello H, editors. C-ORAL-BRASIL: corpus de referência do português brasileiro falado informal I. Belo Horizonte: Editora UFMG; 2012.

20. Raso T, Mello H, Ferrari L. C-ORAL-BRASIL II. To appear.

21. Raso T, de Melo Rocha BNR, Salgado JV, Cruz BF, Mantovani LM, Mello H. The C-ORAL-ESQ project: a corpus for the study of spontaneous speech of individuals with schizophrenia. Lang Resour Eval. 2024;58(3):903–23.

22. Reed BS. Beyond the particular: Prosody and the coordination of actions. Lang Speech. 2012 Mar;55(1):13–34.

23. Rocha B, Raso T, Mello H, Ferrari L, et al. Information structure in the speech of individuals with schizophrenia: Methodology and first analyses from complex structure of corpus based data. CHIMERA Rev Corpus Lenguas Romances Estud Linguísticos. 2022;9:217–42.

24. Schegloff EA. Overlapping talk and the organization of turn-taking for conversation. Lang Soc. 2000;29(1):1–63. doi:10.1017/S0047404500001019.

25. Šturm P, Skarnitzl R, Nechanský T. Prosodic accommodation in face-to-face and telephone dialogues. In: Interspeech; 2021. p. 1444–8.

26. Vinciarelli A, Chatziioannou P, Esposito A. When the words are not everything: the use of laughter, fillers, back-channel, silence, and overlapping speech in phone calls. Front ICT. 2015;2:4.

27. Vital Á, Rocha B. Disfluencies classification: a corpus-driven approach. In: Proceedings of the 2nd International Conference on Data & Digital Humanities: Generative Artificial Intelligence for Text and Multimodal Data; University of Minho (Portugal). In press.