Topic unit detection in spontaneous speech
Measuring reliability using the Kappa statistic

Frederico Amorim Cavalcante°, Tommaso Raso°, Giulia Bossaglia°, Maryualê Mittmann*, Bruno Rocha°

°Universidade Federal de Minas Gerais, *Universidade do Vale do Itajaí

This paper deals with an inter-annotator agreement test involving the identification of the information unit of topic as defined within the framework of the *Language into Act Theory* (L-AcT). Fleiss's kappa statistic was used to measure the agreement among the four annotators who took part in the test. The data used was sampled from C-ORAL-BRASIL II, a spontaneous speech corpus of Brazilian Portuguese. The paper begins by outlining of the theoretical underpinnings of L-AcT, dedicating special attention to aspects directly related to the notion of Topic. Section 2 presents the pilot test and discusses methodological and theoretical issues that were relevant for the design of the protocol that was eventually used in the actual test. Sections 3 and 4 deal with the test, its protocol and results (the kappa coefficient for the general agreement was 0.79, which by usual standards represents a substantial agreement). Section 5 first provides a brief review of a few studies conducted according to other frameworks which have dealt with inter-annotator agreement on the annotation of information structure categories. Finally, the errors observed in the test are analyzed qualitatively.

Keywords: interrater agreement, information structure, topic, spontaneous speech, prosody.

1. Introduction

This paper deals with the interrater agreement on the detection of the information unit of Topic as defined within the framework of *Language into Act Theory* (L-AcT; Cresti 2000; Cresti 2018; Moneglia & Raso 2014; Cavalcante 2020). L-AcT constitutes a pragmatic framework for speech analysis developed on the basis of

naturally occurring data collected from corpora[1]. Some of the core assumptions of this framework are as follows[2].

- The speech flow is segmented by means of terminal and non-terminal prosodic boundaries. The speech sequence between two terminal boundaries, called *terminated sequence* (TS), is pragmatically and prosodically autonomous, and carries at least one illocutionary prosodic unit. TSs can be of two types: utterances and stanzas. While utterances are composed of one single pattern of prosodic units, stanzas are composed of a sequence of juxtaposed sub-patterns, linked to one another by a prosodic signal of continuity. Stanzas are always segmented into more than one prosodic unit by means of non-terminal prosodic boundaries. Utterances that feature non-terminal prosodic boundaries are called *compound utterances*, whereas those that do not are called *simple utterances*, and their single prosodic unit necessarily carries the illocution.
- The illocution constitutes the nucleus of the utterance and of each stanza sub-pattern. No communicative function can be attributed to a speech sequence that lacks a unit conveying the illocutionary force.
- Compound utterances and many stanza sub-patterns are composed of the prosodic unit that conveys the illocutionary force plus one or more non-illocutionary units that can convey different informational functions. There is a general isomorphism between prosodic form and informational function: each prosodic unit carries out an informational function; moreover, each informational function is associated with a particular prosodic form showing characteristic f0, intensity, and duration patterns.
- Information units are defined by their function, prosodic shape, and distribution inside the TS. The Comment is the information unit that carries the illocution. It always features a functional nucleus (Cresti 2011; Raso & Rocha 2017) and its form varies according to the type of illocution that it conveys. L-AcT recognizes five other types of information units which contribute to the building of the semantic content of TSs. These units, as well as the Comment, are called textual information units. In addition, L-AcT recognizes the so-called dialogic information units, a group consisting of

[1] For details about the main corpora compiled in accordance with the principles established by L-AcT, see Cresti (2000), Cresti & Moneglia (2005), and Raso & Mello (2012)

[2] For a comprehensive explanation of the theoretical framework, see the references provided in the above footnote.

approximately seven units – the precise number is yet to be defined – whose functions are similar to those fulfilled by what is referred to as Discourse Markers within other frameworks (Cresti (2000) proposes five dialogic units; Frosali (2008) proposes a sixth one; Raso (2014) shows that dialogic units can be considered Discourse Markers understood in terms of a specific group of information units; Raso & Vieira (2016), Gobbo (2019) and Raso & Ferrari (forthcoming) describe these units on the prosodic level and analyze some of them statistically). While dialogic information units deal with the regulation of the channel and the interaction with the addressee, textual information units are responsible for the semantic content of the utterance. In addition, there is a specific type of dialogic unit, called *discourse connector* (see Cresti and Moneglia 2019), that has a cohesive function. As for the textual units, besides the aforementioned comment, which conveys the illocution, there is the topic, which will be explained below, and also the *parenthetical* (Tucci 2010), a unit that has been recognized within other frameworks as well; the *appendices* of *comment* and of *topic*, which constitute units that, without featuring any functional prosodic prominence, integrate the text of the comment and of the topic; and, finally, the *locutive introducer* (Maia Rocha & Raso 2011), which has the function of introducing meta-illocutions, mainly reported speech. Each type of informun unit features a characteristic prosodic profile and has distributional constraints.

The examples below show different instances of information units. Example (1) shows both simple and compound utterances in context, whereas example (2) shows a stanza[3].

(1) afamdl01 [link to ex1.wav]
*FRE: [16] see /=AUX= the day before yesterday /=TOP= I did ice cream /=COM= right //=AUX=

[3] Examples (1) and (2) were taken from a informationally annotated minicorpus of American English (Cavalcante & Ramos 2016; Cavalcante *et al.* 2018), which was derived from the *Santa Barbara Corpus of Spoken American English* (Du Bois *et al.*, 2000-2005). The alphanumeric codes introducing each example (namely, *afamdl01* and *afammn02*) mean that the examples are taken from the American English minicorpus (hance the initial "a"), from the family section (hence "fam"). The first example is a dialogue (hence "dl"), and the second a monologue (hence "mn"). The digits specify the rank of the texts in the subsection of the minicorpus from which they come (01 and 02, respectively).

*FRE: [17] Balian //=COM=
*RIC: [18] hum hum //=COM=
*FRE: [19] and you gotta pack those in cases //=COM=
*FRE: [20] <and so> /=AUX= like /=AUX= I didn't put that down on my production <card> //=COM=
*RIC: [21] <right> //=COM=
*RIC: [22] <how> many cases you packed //=COM=
*FRE: [23] I don't know /=COM= man //=AUX=
*FRE: [24] I packed two pallets /=COM= you know //=AUX=

The example above shows a sequence of utterances taken from a dialogue[4], and, as already mentioned, it contains both simple and compound utterances. The utterances numeber [17], [18], [19], [21], [22], and [23] are simple, while the others are compound utterances, each of which illustrating different types of configuration.

(2) afammn02 [link to ex2.*wav*]
*ALN: [10] &he /=TMT= flew down to Mexico City /=COB= &he /=TMT= we &c [/1]=SCA= think of the name of my hotel /=COB= which wouldn't mean anything now /=PAR= but we ended up in a /=SCA= fabulous hotel /=COB= &he /=TMT= first night /=TOP= we were very unhappy with our rooms /=COB= we got down there //=COM=

Example (2) above shows a stanza[5]. This sequence of information units features a terminal prosodic boundary only at the end, but it contains five illocutionary units. The final illocutionary unit is tagged as COM to show that it is marked by a terminal break, while the other illocutionary units are tagged as COB (bound

[4] The three-letter tags in the utterances stand for: comment (COM), i.e. the illocutionary unit, which is mandatory for a sequence to be interpreted as an utterance; dialogic auxiliary (AUX), an umbrella tag that represents all of the dialogic units; and topic (TOP). For details about TOP, see Section 1.1 below.

[5] The three-letter tags in example (2) stand for: time taking unit (TMT) also called filled pause, i.e. a strategy for keeping the turn while elaborating the message; bound comment (COB), a processual illocutionary unit inside a stanza, marked by a prosodic signal of continuation instead of the terminal signal characteristic of a regular comment (COM); parenthetical (PAR); and scanned unit (SCA), an intonation unit constituting an information unit that is made up of more than one intonation unit (see Section 2.2).

comments) to show that they are marked by a prosodic signal of continuity. Thus, five juxtaposed illocutionaty subpatterns can be seen in this stanza. While the first, the third and the last subpatterns contain only the illocutionary information unit, since the other intonation units have no informational value, the second and the fourth subpatterns contain different non-illocutionary information units: the second subpattern has a parenthetical at its end, and the fourth hass a topic unit.

The next section will deal with the information unit with which this paper is directly concerned.

1.1 The information unit of Topic

As stated earlier, this paper focuses on the information unit of topic (TOP), which has the function of supplying a cognitive domain for the interpretation of the illocution. When an utterance does not have a TOP, the cognitive domain for the application of the illocutionary force is established on the basis of the linguistic or situational context. TOP always occurs to the left of the illocutionary unit and is realized by prosodic units with specific melodic and durational patterns (Raso *et al.* 2017; Cavalcante 2020).

As can be seen, the definition of TOP according to L-AcT differs from more traditional definitions found in the information structure literature, particularly because L-AcT regards TOP in terms of *pragmatic aboutness* (a domain for the illocution), whereas in other approaches the notion associated the term topic is usually established in terms of *semantic aboutness* (a domain for the predication). To give a few examples, according to Li and Thompson (1976) and Chafe (1976), the topic of a sentence specifies the framework within which the predication holds. Krifka (2008), on the other hand, defines topic somewhat loosely as that which the sentence is about, a semantic notion that does not take the illocutionary dimension into consideration. For L-AcT, on the other hand, TOP establishes the domain for the interpretation of the speech act – see Cavalcante (2020) for more on how L-AcT differs from other approaches.

Another important aspect of the definition of TOP by L-AcT is that, in principle, there is no morphosyntactic constraint for the locutive content of the unit, which nonetheless must correspond to a cognitive domain of identification. Although many TOPs correspond to noun phrases on the morphosyntactic level, they can also correspond to verb phrases, prepositional phrases, adverb phrases and even adjective phrases.

The semantic requirement that TOP must correspond to a domain of identification arises from the impossibility for a functional word or a negative NP to function as TOP, unless they are used as a citation, as the example below shows.

(3) "for" /TOP is one of the most common prepositions in English //
"Nobody" / TOP is the name Ulysses gave to himself when asked by Polyphemus //

Likewise, an NP with non-specific interpretation cannot function as TOP. In the following example, the NP in TOP must receive a generic interpretation. In other words, this NP has to be interpreted as referring to a class in order for it to be able to properly function as TOP:

(4) A priest / TOP should be an ethical person //

Thus, both the prosodic and the semantic characteristics that allow to individuate a cognitive domain of identification must be present for a unit to be recognized as the topic of an utterance.

The distributional requirement that TOP must occur to the left of the illocutionary unit is supported by the fact that appendices cannot provide a cognitive domain for the illocution, and thus cannot be considered postponed topics. In fact, the semantic content of appendices (units that always exhibit flat or falling *f0* profile) must always be given/known, as these units cannot instantiate a new cognitive domain. On the other hand, TOP does instantiate new cognitive domains, changing the cognitive focus of the utterance that comes before it to a new one[6].

In addition, it is important to mention that, according to L-AcT, syntax is a level that is substantially contingent on information structure interpretation (Cresti 2014). So, a constituent potentially analyzable as a syntactic subject can in fact constitute the morphosyntactic makeup of a TOP unit. In other words, there are TOP units that are liable to be interpreted as a subject if their prosodic features – i.e. non-terminal break and prosodic form – are overlooked. A constituent functioning as TOP is not cognitively construed the way a constituent functioning as subject is, after all, to behave as a cognitive domain for an illocution, which is what a TOP unit does, differs from behaving as the syntactic subject within a predicative structure.

1.1.1 *The prosodic forms of TOP*

As already mentioned, TOP is realized by specific prosodic forms. Studies conducted within the L-AcT framework have identified and statistically validated

[6] For a comprehensive description of the semantic aspects of TOP, see Cresti (2000) and Firenzuoli (2003).

three different prosodic forms of TOP (see Raso et al. 2017; Cavalcante 2020). Each prosodic form is composed of a small number of syllables, called the *nucleus*, which are responsible for imparting the function of the unit. Non-nuclear syllables, if there are any, play a role on the semantic and syntactic levels, as they have no informational function. They are responsible for giving a linguistic form to the locutive content of the unit. The nucleus of all three prosodic forms, besides having specific melodic shape, is characterized by syllable lengthening and generally higher intensity values. The three types of prosodic forms of TOP are described in more detail below.

- Type 1 form has a rise-fall *f0* movement that begins on the last stressed syllable of the unit and, there being post-stressed syllables, continues onto them.

(5) once I get my experience /=TOP= I 'll be up there too / in the top four salesman //
afamdl01_ 80

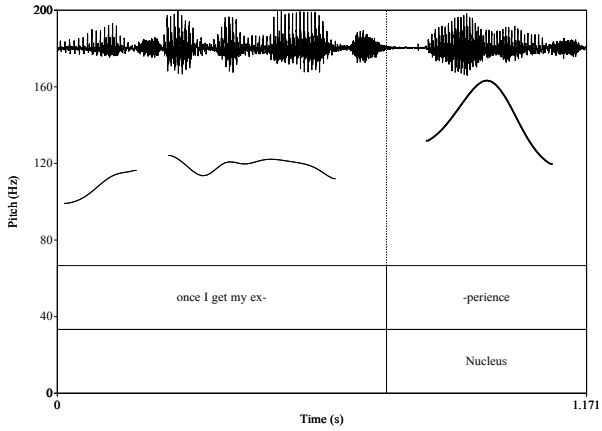

Figure 1: Waveform and *f0* curve of the Type 1 TOP shown in example (5) above. The portions corresponding to the nucleus is indicated in the first tier of the grid.

- Type 2 form has a rising *f0* movement that begins on the last stressed syllable of the unit, continuing onto the post-stressed ones if there are any present.

(6) but in a sense /=TOP= I need a [/1][7] some type of steady income // afamdl01_67

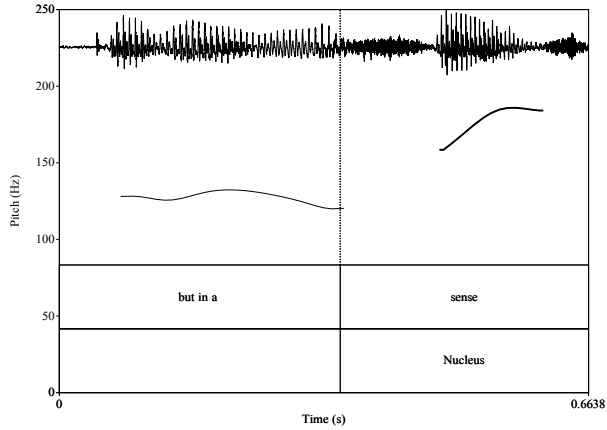

Figure 2: Waveform and *f0* curve of the Type 2 TOP shown in example (6) above. The portions corresponding to the nucleus is indicated in the first tier of the grid.

- Type 3 form holds two often discontinuous semi-nuclei. The first one features high to extra-high f0 values, while the second one features lower f0 values. Sometimes this prosodic from features non-nuclear syllables following the second semi-nucleus. These syllables are referred to as *codas* and correspond to the part in italics in example (7) below.

(7) when Mary tells me to get a sleep *over the weekend* /[TOP] you know I need to get sleep over the weekend // afamcv04_138

[7] Convention used to indicate that one word (in this case the article *a*) has been retracted. If more words are retracted, the digit following the forward slash will specify the number of words retracted.

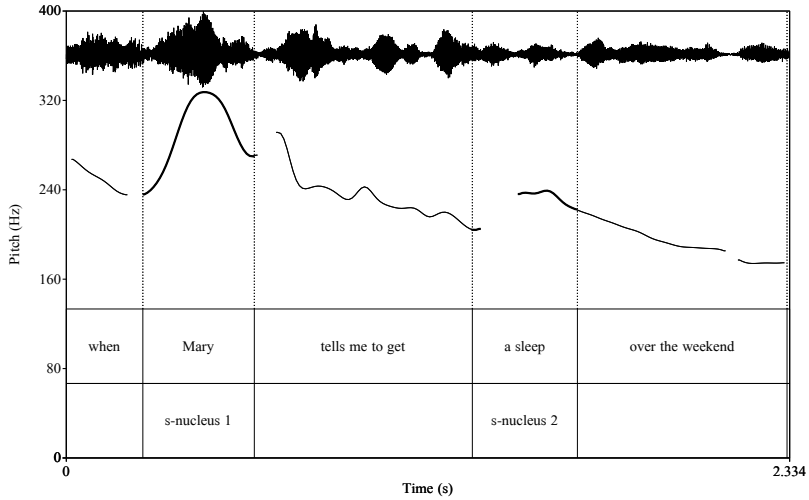

Figure 3: Waveform and *f0* curve of the Type 3 TOP shown in example (7) above. The portions corresponding to the semi-nuclei are indicated in the first tier of the grid.

Codas are more frequently found in American English (AE) than in the Romance languages that have been studied so far. In addition, the frequency of a form in a language may vary considerably, which is likely due to the rhythmic structure of the language in question (see Raso *et al.* 2017 and Cavalcante 2020).

The goal of this study was to establish how reliably can TOP units be detected in spontaneous speech data by conducting an interrater agreement test with four participants and using spontaneous speech data from C-ORAL-BRASIL II (Raso *et al.* forthcoming and Bossaglia & Ferrari 2019).

C-ORAL-BRASIL II constitutes a collection of corpora of Brazilian Portuguese speech recorded in a number of different contexts, documenting formal face-to-face exchanges, media and telephone interactions[8]. Together with C-ORAL-BRASIL I, it constitute a corpus comparable with the individual components of the C-ORAL-ROM multilanguage corpora (Cresti & Moneglia 2005), which document Italian, Spanish, European Portuguese, and French.

[8] The formal part of the natural context corpus contains 74 texts totaling 121,396 words; the media corpus contains 101 texts totaling 139,396 words, and the telephonic corpus contains 79 texts and 31,308 words. In total, the C-ORAL-BRASIL II contains 289,921 words.

In order to assess the reliability of TOP unit detection, we devised a test in which we used Fleiss's kappa statistic (Fleiss 1971; Davies & Fleiss 1982) to measure the level of agreement among four experienced annotators who had received training on information-structure annotation according to the L-AcT framework. We began by conducting a pilot test designed to determine the best possible protocol for the actual test. But before presenting it, we will discuss some other reliability tests conducted within other frameworks for the study of information structure.

1.2 Other agreement tests

Quantitative analyses of information structure annotation are not as frequent as desirable (see Lüdeling *et al.* 2016 and Ritz *et al.* 2008). Furthermore, the studies that do seek to measure the degree of annotation reliability in more objective terms sometimes show their results only in percentages, thus failing to account for agreement due to chance (e.g. Vaselá *et al.* 2004). The agreement strength reported in these studies vary considerably, depending on the information structure notion considered – topic, focus, information status etc. – and the methods employed in the experiments, particularly the level of complexity of the annotation task and the profile of the participants. Although it is hard to draw direct comparisons among different frameworks, given the often-profound theoretical differences that set them apart, we believe that some studies focusing on agreement on information structure annotation carried out within frameworks other than L-AcT merit some attention.

Ritz et al. (2008) conducted quantitative and qualitative evaluations of the LISA scheme (Dipper *et al.* 2007) using 49 texts in German, both written and transcribed (which means oral texts are deprived of all the information conveyed by the acoustic signal), totaling 1,940 tokens, 515 NPs and 271 sentences. Their data consisted of (i) map task dialogues, (ii) question-answer pairs elicited with QUIS[9], and (iii) online newspaper comments from the *Potsdam Commentary Corpus* (PCC; Stede 2004). The authors report kappa coefficients expressing the degree of agreement on the annotation of different information structure categories by two undergraduate students who underwent an intensive and short training period (half a day).

The categories were grouped into three classes: information status, focus, and topic. *Information status* encompassed some notions related to givenness and ac-

[9] A questionnaire for eliciting data for information structure research (Skopeteas *et al.* 2006).

cessibility, whereas *focus* encompassed contrast and new information. The category called *topic* was associated with the notion of *semantic aboutness topic* (i.e., an expression that expresses the entity that the sentence is about), and *frame setter* (i.e. an expression that delimits the domain to which the main predication applies). The authors used the kappa statistic to compute results for each of the three classes.

Regarding the results specifically for topic, although the term is associated with two different functions (aboutness topic and frame setters), the results are presented for the class as a whole. The level of agreement varies from $k = 0.44$ to $k = 0.91$, with higher results for sentences from question/answer texts and when the annotation was carried only on noun phrases. Question/answer texts yielded much better results likely because, when annotating them, one may use the question to reliably establish the aboutness topic of the sentence that answers it. The same cannot be said of the two other texts types, namely, map task dialogues and online newspaper comments. The authors noted ambiguous examples, particularly from the PCC sample, in which a sentence exhibited two aboutness-topic candidates. This may be taken as an indication that the notion of semantic aboutness topic need to be reviewed, considering that it is often hard to apply it in a consistent way, especially in natural texts.

It must be mentioned that Ritz *et al*. makes no actual distinction between written and spoken data. In practice, the latter is treated as the former, since only transcriptions appear to have been considered during the annotation. According to the framework guiding the present study, annotating information structure in spontaneous speech cannot be done without recourse to prosody. In the absence of the prosodic input, one cannot even establish the syntactic relationships among a string of transcribed words, let alone determine their informational function.

Paggio (2006) applies a binary annotation scheme based on Lambrecht (1994) in an experiment with two raters and spoken data in Danish taken from the DanPASS corpus (Grønnum 2006), which is not a real spontaneous speech corpus, since its main goal is for phonetic studies. In this case, prosodic cues were taken into account, but the definition of topic and focus, the two categories that were annotated, does not take the illocutionary level into account. The degree of agreement reported ranges from $k = 0.7$ to $k = 0.8$, and the disagreement are mostly due to the identification of focus, particularly the marking of its left margin, which annotators tend to disagree, for example, on whether it includes the nucleus of the VP or only its internal arguments. Besides this, little is said about the disagreements observed, so it is hard to say much about these results. There is no information about the agreement on focus and topic considered separately.

More recently, Cook and Bildhauer (2013) devised two experiments in which the interrater agreement was measured on the distinction of thetic (i.e., topicless) sentences from categorical (i.e., topic-comment) ones and the identification of semantic aboutness topic. The first experiment can be regarded as pilot test, and it was used to improve the guidelines for the second experiment.

The sentences used in the experiments come from *The Mannheim German Reference Corpus* (DeReKo; Kupietz *et al.* 2009), which contains naturally occurring written texts in German, and the TüBa-D/Z treebank (Beck 2012), which feature German newspaper texts. The first experiment was carried out with two experts (the authors of the paper) and the second with four individuals with some background in linguistics (no more than 4.5 years of training) but no prior experience with annotation of information functions.

In Cook and Bildhauer's (2013) first experiment, the task involved looking at detached sentences, all of which containing one of four previously defined verbs, and deciding whether they were thetic or not and then determining whether preselected phrases (e.g., subject-NPs, objects-NPs, and adverbs such as *here*, *then* and the like) were functioning as aboutness topics or not. The kappa coefficient computed in the first experiment for the thetic versus categorical distinction was low (considering all sentences, $k = 0.44$, and always less than when sentences were sorted according to verb). Regarding the identification of aboutness topics, the coefficients ranged from $k = 0.19$ and $k = 0.57$, depending on the verb considered. The causes of this variation, however, are hard to pinpoint, but, to a considerable extent, they seem to be due to the fuzziness associated with the definition of categories in question, which makes it hard for them to be applied consistently.

To give an indication of the confusion associated with these categories – and hence with the criteria for identifying them –, the authors even speculate about accent placement while analyzing ambiguous cases of aboutness topic in their data, which, again, is entirely made up of written material (see Cook and Bildhauer 2013, p. 127).

For the second experiment, the four participants were given texts (rather than sentences) with NPs and PPs marked beforehand so that they knew exactly which phrases could potentially be functioning as aboutness topics. Its poor result ($k = 0.447$) for agreement on aboutness topic identification is unsurprising, given that the guidelines used in it constituted a version of those used in the first experiment, thus being tainted by standards that are unclear and, sometimes, even alien to the type of data that was being used in the experiments.

In discussing the results, Cook and Bildhauer (2013) point out that the two experiments yielded disappointing results, and they raise the possibility that the concept of aboutness topic may not be amenable to operationalization, since it is

often hard to even determine whether a naturally occurring sentence has an aboutness topic in the first place or whether it is a thetic sentence. They conclude by suggesting that it may be impossible to devise language-independent guidelines for information structure annotation.

In the next section, we discuss the pilot test we conducted in preparation to our reliability test on the detection of TOP units in spontaneous speech.

2. The pilot test

For the pilot test, five subjects[10] were selected among the most experienced researchers from the *Laboratory of Empirical and Experimental Linguistic Studies* (LEEL[11]), where the project responsible for the compilation of the C-ORAL-BRASIL[12] corpora is developed (Raso *et al.* forthcoming, Raso & Mello 2012). The selected participants not only are familiar with the L-AcT approach but also have extensive experience with information structure annotation at the time of the test.

The sample used in the pilot test consisted of 50 compound, non-interrupted terminated sequences (TSs) from the C-ORAL-BRASIL II corpora (Raso *et al.* forthcoming). This means that the sample contained only fully accomplished TSs – i.e., marked with a terminal prosodic break – composed of more than one prosodic/information units. All TSs were taken from the C-ORAL-BRASIL II *Natural Context* section, which contains formal face-to-face interactions including lectures, political speeches, court hearings, and so forth. The *Praat* software (Boersma & Weenick 2019) was chosen for the annotation task so that annotators could not only listen to the utterances but also have access to visual cues such as f0 curves and spectrograms.

The TSs were organized in a text files to be used by the participants and a protocol was set up which established that participants should carry out the annotation according to the three following steps:

a. identification of illocutionary units;
b. identification of TOP units;

[10] The participants correspond to the authors of this paper.
[11] Portuguese: *Laboratório de Estudos Empíricos e Experimentais da* Linguagem; LEEL website: http://www.letras.ufmg.br/leel/.
[12] C-ORAL-BRASIL website: http://www.c-oral-brasil.org

c. identification of the remaining prosodic units without distinguishing their informational functions.

The reason for identifying illocutionary units prior to identifying TOP units is basically that the illocution constitutes the nucleus of the utterance and its recognition is essential both to recognize a given sequence as a TS and to understand the relationships established between the illocutionary and the non-illocutionay information units. Thus the identification of a TOP is contingent upon the identification of the "actional" (illocutionary) nucleus of a TS or one of its sub-patterns.

In addition to annotating the prosodic units, the participants were asked to report any problems experienced during the test as well as mistakes observed in the transcriptions and prosodic segmentation, which they were not allowed to modify.

2.1 Results of the pilot test

The 50 utterances of the sample added up to 260 prosodic units. The number of TOPs identified by the annotators ranged from 17 to 35 (M = 23.6, SD = 8.0). Thus, about 9% of the prosodic units in the sample were tagged TOP, which is consistent with prevalence estimates for the unit reported in previous studies (Mittmann 2012; Cavalcante 2015).

In order to measure the extent of the agreement among the annotators, we used the Fleiss's kappa statistic[13], which measures the reliability of the agreement among more than two raters. The main advantage of computing kappa coefficients instead of simply reporting agreement proportions is that the kappa statistic takes chance agreement into account, thus allowing for more reliable conclusions to be drawn.

The kappa coefficient expressing the overall level of agreement in the pilot test was k = 0.51. According to the benchmarks proposed in Landis and Koch (1977: 165), this coefficient expresses moderate agreement, as seen in Table 1.

[13] For details about the kappa statistic, see Fleiss (1971) and Davies & Fleiss (1982).

Table 1. Kappa coefficients and degree of agreement.

Kappa statistic	Strength of agreement
< 0	Poor
0.0 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost perfect

We also computed *k* considering four annotators at a time so as to check whether any of them were particularly influencing the strength of the agreement (see Table 2 below). Note that in Table 2 whenever the E's annotation is being considered, the agreement is considerably lower.

Discussing reliability tests similar to the one reported here, Krippendorff (2004a: 211) points out that, to be considered reliable, a procedure must respond to the same phenomenon in a uniform way, no matter the circumstances in which the procedure is applied. Therefore, an agreement test with different annotators can be regarded as an attempt to determine how sure one can be that a given set of criteria is being consistently followed by the annotators, so as to be able to infer whether the implementation of such criteria can be replicated or not. To be sure, tests of this kind do not allow us to conclude anything about the nature of the category being annotated, meaning that these tests have nothing to say about the validity of the phenomenon being annotated. In line with that, our objective in this paper is to assess the extent to which the specifications provided by L-AcT for the identification of TOP may lead to trustworthy results. Landis and Koch's (1977) cutoff values shown above should be regarded as indicative of the consistency of the research procedure in question, and they cannot be used to make a case for the validity of the category under examination. Finally, these cutoffs should never be applied uncritically, given that their interprepretation is a delicate matter that depends on the methods and questions of a given study. For a critical appreciation of coefficients that measure inter-annotator agreement, see Krippendorff (2004a, 2004b).

It is worth mentioning that the annotator in question had spent a very long time – approximately five years – dealing with information structure annotation only occasionally, which prevented her from being completely up to date on the developments related to the task at the time of the pilot test. So, we decided that this person, instead of taking part in the actual test as an annotator, would take care of the transcriptions to be used in it by revising their prosodic segmentation.

No similarly expressive change in *k* was observed upon removing any of the other four annotators.

Table 2. Kappa coefficients computed taking four annotators at a time.

Annotators	**K statistic**
B, C, D, E	0.45
A, C, D, E	0.46
A, B, D, E	0.49
A, B, C, E	0.50
A, B, C, D	0.61

Full disagreements (i.e., only one annotator classified a given unit as TOP) occurred in 22 cases, full agreements (i.e., all annotators classified a given unit as TOP) in 6 cases, and partial agreements (i.e., at least two annotators identified the unit as TOP) in 23 cases. Thus, considering both full and partial agreements together, there were 29 cases of agreement.

2.2 Some observations based on the pilot test

The analysis of the results, which were discussed during meetings in which annotators A, B, C, and D participated, suggested that some of the disagreements observed were caused by inattention on the part of the annotator as to the semantic aspect of the definition of TOP.

This occurred in cases where a unit with prosodic features resembling those of TOP was realized by a piece of locutive content that could not be supplying a cognitive domain for the interpretation of the illocution. The Portuguese expression *então* 'then/so' in the utterance below is a case in point. While the melodic curve with which it is realized suggests it constitutes a TOP, from the semantic perspective – and when considered out of context – it may either convey a temporal meaning, thus being able to constitute a cognitive domain, or carry out a dialogic role, in which case it cannot be TOP.

(8) *então* / o que que a gente vai fazer aqui nesses dois dias //
'so / what is it that we're going to do here during these two days //'
bnatte05_48

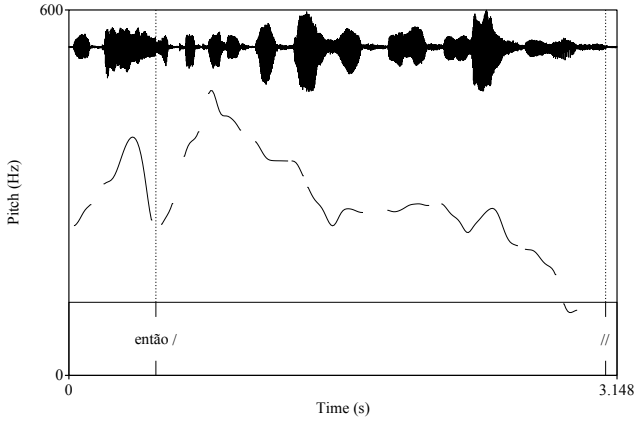

Figure 4. F0 curve of the utterance in example (8).

In fact, the expression *então* in the utterance above is the locutive content to a *dialogic unit*. As already mentioned, in order for a unit to be TOP, its locutive content must be referential. Thus, *então* would have to mean something along the lines of "at that moment" in order to qualify as TOP. This requirement, however, is not met, as the English translation beneath the utterance transcription suggests. Despite that, one of the participants of the pilot test, motivated by the rise-fall movement of the melodic curve seen in Figure 4[14], tagged it TOP.

The analysis of the results also pointed to another source of inconsistency, namely, the reduced number of syllables in certain prosodic units. Although units with a small number of syllables can indeed be TOP units, they are likely to play other information functions. When a TOP-candidate expression has one or two syllables only, unless the prosodic form it shows is particularly clear or else the likelihood of syntactic compositionality is ruled out, it may be quite hard for one to determine whether the expression is composing a TOP unit or not, for it may well be the case that the expression is composing a *scanning unit* (SCA) carrying the subject of a sentence in the illocutionary unit.

SCA units constitute the only case in which the isomorphism between prosodic and information units does not hold. The realization of a piece of locutive content as a SCA unit may be accounted for by issues involving disfluency, emphasis, and articulation problems that can occur while packaging large strings of

[14] For details about the prosodic features of TOP, see Raso *et al.* (2017) and Cavalcante (2020).

words into just one prosodic envelope. The prosodic boundary following a SCA unit is not interpreted as signal that ends syntactic compositionality, which enables one single information unit to be realized by more than one prosodic unit. The prosodic features of a scanned information unit always sit in the last prosodic unit. Example (9) shows two indisputable cases of SCA units.

(9) isso aqui são dados da Capes / são dados muito interessantes // que são /SCA *os acessos* /SCA ao portal da Capes //
'this over here is data from Capes / they're very interesting data // which are / the visits / to Capes website //'
bnatco08_99-100

Controversies emerged in cases such as shown in example (10), where *eles* 'they-masculine' could in principle be interpreted either as a scanned subject or a cognitive domain for the illocution. The prosodic unit holding the pronoun *eles* in the example below was tagged TOP by one of the annotators, while the others labeled it SCA. A careful examination of the audio file reveals no prosodic feature that can be associated with the information function of TOP, and we can thus conclude that *eles* is compositional with the locutive content of the subsequent unit, which is the one that in fact carries the prosodic signal of the informational value the sequence.

(10) *eles* /SCA não só propõem essa mudança / (...)
'they / not only propose this change / (…)'
benatpd09_165

We also observed problems regarding the distinction between verbal TOPs and illocutions in stanza sub-patterns. A verbal TOP has the ability to establish a cognitive domain by virtue of the temporal, causal, concessive or some other referential meaning that it conveys. The prosodic features of verbal TOPs and these illocutions are sometimes similar, given that the prosodic form of a TOP may suggest the continuity signal that characterize illocutions in a stanza[15]. However, the semantic interpretations of TOP units and illocutionary units are clearly different. This type of confusion was taken into consideration during the preparation for the actual test.

[15] For a discussion about the prosodic patterns of TOP and COB, see Cavalcante (2020).

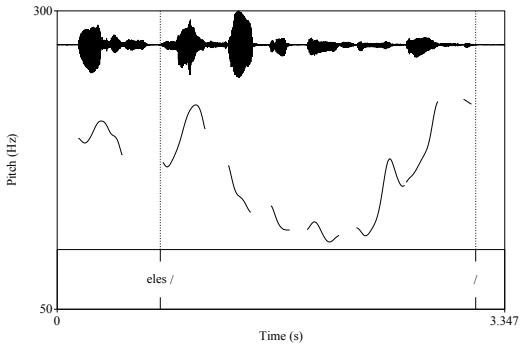

Figure 5. F0 curve of the utterance in example (10).

Finally, we noted that the lack of the preceding context – the TSs used in the pilot test were presented without the aid of surrounding TSs – sometimes created problems, particularly when the referential interpretation of a TOP-candidate expression was at stake. In addition, it occasionally happened that an annotator did not agree with the placement of the terminal break in the transcription, a problem that could easily be solved if they had been provided with a couple of extra TSs immediately following the target one.

Based on the observations outlined above, we set up a protocol for the actual test, which is discussed in the following section.

3. The agreement test

In this section we present details about the agreement test conducted to determine the reliability of the criteria established by L-AcT for the identification of TOP units in spontaneous speech data.

3.1 The protocol

As in the pilot test, we used *Praat* (Boersma & Weenick 2019) for the realization of the test. The participants were instructed to disable the *Pitch visualization* function in the editor window of the application and to base their annotation entirely on auditory cues and semantic/pragmatic analysis, so as to avoid biases that *f0* curves may cause and also to approximate the natural conditions of speech interpretation.

The annotation steps of the pilot test were also kept. So, for each TS, the annotator should first identify the illocutionary unit, then check for TOP units, and finally label remaining units without specifying their informational function.

The tagset to be used in the annotation was specified in the protocol set up based on the pilot test and is listed below:

- COM for illocutionary units within regular utterances;
- CMM (Multiple Comment) for illocutions in a Compositional illoctionary pattern, which consists of a sequence of two or a few more illocutions that are prosodically patterned in a holistic manner thereby conveying a rhetorical effect of comparison, reinforcement, listings or others;
- COB for illocutionary units in a stanza sub-pattern;
- TOP for topic units;
- NTP for all other kinds of information structure.

The protocol also contained instructions designed specifically to guide the identification of TOP units, meant to ensure that participants would be using the same criteria. The guidelines contained the following information:

a. TOP is always referential, as it must constitute a cognitive domain of identification for the illocution. The domain can be individual, conceptual, temporal etc.
b. Units made up of short pieces of locutive content (e.g., stand-alone pronouns) characterized by a relatively weak phonetic realization should be very carefully considered. Such units should never be labeled TOP in the absence of a clear prosodic profile leading to the incontrovertible interpretation of a cognitive domain for the illocution and ruling out the possibility of compositional relationship of a SCA unit.
c. Prosody is an important aspect, but the semantic criterium, which establishes the referential nature of TOP, must always be taken into consideration. In other words, in order to qualify as TOP, a unit has to pass both the prosodic and the semantic tests. This applies also to cases which may be confounded with COBs arising due to a combination of syntactic and prosodic factors when VPs are realized featuring a prosodic signal of continuity.
d. As a corollary of (c), the locutive content of units constituted by stand-alone expressions such as *aí* (‘there’, ‘so’, or ‘then’) and *então* (‘so’or ‘then’) must carry out a clear referential function – which can be of either temporal or spatial nature – in order for these units to be labeled TOP. Otherwise, they should be considered dialogic and thus labeled NTP. It goes without saying

that the prosodic features must also be clearly present, and this applies to all cases where the syllabic extension of the unit is particularly reduced.

e. If a prosodic break is present in the transcription but the annotator deems it either misplaced or absent, NTP should be used.

f. A prosodic unit occurring after TOP and showing a clear prosodic contrast with its neighboring units is unlikely to be another TOP unit, even if its final syllables exhibit a rising *f0* movement.

The cautionary instruction in (f) merits further clarification. The textual-unit group mentioned in Section 1 includes an information unit called *Parenthetical* (PAR), which has the function of providing metalinguistic and modal information meant to aid the addressee in the interpretation of the TS or a part thereof. PAR is marked with prosodic features signaling that it pertains to a level that is distinct from the level of the TS. Example (7) below shows a TS containing a PAR unit with melodic curve that begins with a steep descending movement that eventually turns into a rising curve towards the end of the unit (see Figure 6). The rising portion of the curve resumes the *f0* values to a level that is compatible with the hierarchical level of the TS. When a phenomenon such as this happens, an inattentive annotator may mistakenly consider such a unit to be TOP, given that Type 2 and Type 3 prosodic forms (see Section 1.1 above and Cavalcante 2020) may feature a rising movement at the end. The warning in (f) is meant to prevent this type of mistake.

(11) as mulheres /^{TOP} pelo que eu vi /^{PAR} torcem pelo Fred //^{COM}
the women / as far as I could see / are keeping their fingers crossed for Fred //
bmedts07_202

In examples like the one shown above, a TOP-COM sequence remains perfectly interpretable and coherent from both the perceptual and semantic standpoints if the PAR unit is removed from the audio file (see ex11a.wav). That does not occur if the TOP unit is removed (see ex11b.wav), leaving only the PAR and COM units. This is because PAR is unable to supply a domain of identification for the interpretation of the illocution.

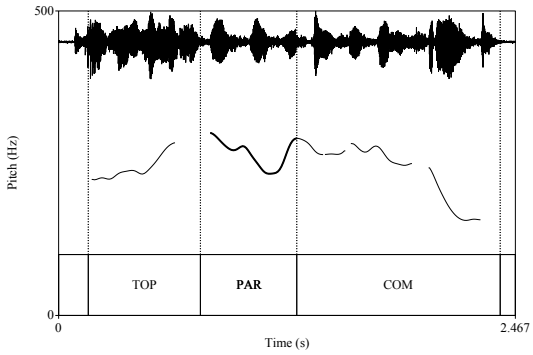

Figure 6. f0 curve of th utterance shown in example (11). The PAR unit corresponds to the thicker portion of the curve.

PAR units can at times occur as relatively long sequences constituting sub-units in common arrangement. One or more of such sub-units may occasionally appear to be functioning as a TOP unit establishing an aboutness relationship with another unit inside the same arrangement of PARs that looks as if carrying illocutionary force. PAR units, however, do not share the same *hic et nunc* or prosodic level with TSs that hold them. These constitute complex cases that cannot be explored in this paper, but they are mentioned here because the participants were asked to pay attention to them in order to avoid labeling TOP a unit inside a long sequence of PARs.

3.2 The data used in the test

The sample that was used in the test consisted of 100 compound TSs randomly sampled from the three parts of C-ORAL-BRASIL II, namely the Natural context section, which documents the formal, face-to-face interactions, the Media section, which encompasses exchanges recorded from TV and radio shows, and the Telephonic section, which documents short telephone conversations pertaining to both the family/private and public domains (see Raso 2012 and Mello 2014).

We decided to sample 20% of the TSs from the Telephonic section, since it holds only about 15% of the compound TSs in C-ORAL-BRASIL II. The remaining 80% of TSs were sampled in equal proportions from the Media and Natural Context sections, which respectively contain 48% and 37% of TSs in the corpus. Thus, a stratified sampling method was used so as to create a sample that mirrored the overall structure of the C-ORAL-BRASIL II.

We began the process of obtaining the TSs by excluding the texts which contained less than five utterances, which was a measure important for texts in the Telephonic section, which can be rather short – in some sporadic cases, they can have even less than five utterances. The reason for excluding such short texts is that we had decided that it was important to provide each *target TS* – i.e., the one that should be annotated – accompanied by four surrounding utterances, two immediately before and two immediately after it. This was meant to help the annotators in two foreseeable circumstances:

- cases in which the reference of a TOP-candidate expression cannot be safely decided by examining only the TS in which it occurs,
- cases in which the annotator may come to disagree as to the placement of the terminal break in the transcription.

Having excluded the short texts, we randomly selected 40 texts from Media, 40 from Natural context, and 20 from Telephonic. The 100 target TSs used in the test came from these texts, each of which providing only one target TS.

The random selection of TSs was carried out using *R* (R Core Team 2019). Once they had been selected and properly organized in spreadsheets, we proceeded to obtain their respective audio files. In this phase, we used *C-ORAL-Search*[16], a very useful *Praat* script that allows the automatic retrieval of an individual TS's sound source from any text of C-ORAL-BRASIL and from most texts of C-ORAL-ROM, depending on the encoding of the XML file of the text. The audio excerpts corresponding to the TSs of our test were saved in individual WAV files and named as shown in Table 3 below.

Before being sent to participants, the material was carefully revised by the pilot test participant that would not take part in actual test in order to make sure that segmentation and transcription faults would be reduced to a minimum.

After the revision, the average number of prosodic units in each target TS was 5.41, with a range of 2 two 24 prosodic units. But most TSs – precisely 71 – had five or less prosodic units. In total, the sample contained 541 prosodic units. Thus, the material exhibited varied degrees of prosodic and informational complexity.

Participants were sent the TSs organized in an XLSX file containing three spreadsheets, one for each section of the corpus. The target utterances were highlighted, as seen in Table 3 below, in order to better distinguish them from their neighboring TSs.

[16] The script is available at https://vieiramarcelo.wordpress.com/praat-scripts/c-oral-search/.

Table 3. Spreadsheet excerpt with TSs and identification of audio files sent to participants.

Utterances	Audio files
*ROD: [14] hhh eu nem sei onde que é na /1 no /1 nos States aí //	btelpv30_14.wav
[15] onde que é //	btelpv30_15.wav
***RAF: [16] no / Texas //**	**btelpv30_16target.wav**
*ROD: [17] ah / tá //	btelpv30_17.wav
*RAF: [18] tá hhh //	btelpv30_18.wav

We took measures to ensure that the utterances used in the actual test had not been used in the pilot test. The participants were sent the protocol via e-mail, along with the files that they should use, and were asked to complete and revise the annotation within the following 30 days. No discussion among participants was allowed.

The participants in the test correspond to the four annotators with best results in the pilot test. As already said, participant E of the pilot test was tasked with revising the prosodic segmentation of the material for the actual test. Annotators have to undergo extensive training before participating in a test of this kind, so it was impossible to have annotators other than the ones from the pilot study.

3.3 Preparing the data for computing kappa coefficients

The material turned in by participants consisted of the annotated target TSs in the XLSX file they had been sent. The tags for computing the kappa coefficients were extracted and organized as seen in Table 4 below. Besides, the illocutionary and non-illocutionary tags (i.e., COM, CMM, COB and NTP) were replaced with "XXX" (see Table 4), as we were particularly interested in measuring the agreement on the identification of TOP units without differentiating the other units. The agreement was also measured considering the tags as they were sent by the participants.

Table 4. Tags extracted to run the agreement test

Tagged utterance	Tags
[19] a essa altura /=TOP= já dá para ver claramente os sinais da febre do Glaciar Exploradores //=COM=	TOP XXX
[81] a tecnologia /=TOP= chegou pra ficar /=COB= ninguém tem dúvida //=COM=	TOP XXX XXX

Once extracted, the tags of all the four participants were organized in data frames and the kappa coefficients were computed considering the tags of each prosodic unit at a time. The preparation of the data was done using *R* and computations were done using the *kappam.fleiss* function, available in the *irr Package* (Gamer & Lemon 2012).

4. Results

By providing the annotators with the utterances accompanied by their surrounding context, the guidelines containing specifications as to what constitutes a TOP unit and instructions on how to detect it, and barring the visualization of f0 curves, we achieved a high level of interrater agreement that resulted in a kappa coefficient $k = 0.79$. According to the benchmarks by Landis and Koch (1977, p. 165) shown in Table 1 above, this represents a level of substantial agreement. Table 5 below suggests that no annotator in particular was causing the agreement level to change in a relevant way.

Table 5. *k* coefficients computed considering three annotators at a time.

Annotators	***k***
B, C, D	0.81
A, C, D	0.79
A, B, D	0.78
A, B, C	0.78

The number of TOPs identified by each annotator ranged from 48 to 55 ($M = 52$, $SD = 3.2$). Full agreement was reached in 36 cases. In 11 cases, three annotators agreed, and in 5 cases only two did. The number of TOPs that each annotator detected and that was not detected by any other annotator ranged from 2 to 8 ($M = 5.25$, $SD = 2.5$). In total, 15 TOPs were identified by one annotator alone.

Considering the utterances from Natural Context, Media, and Telephonic separately, the agreement levels were also substantial. In the Media section, whose 40 utterances added up to 234 prosodic units, the overall agreement achieved was $k = 0.80$, and in the Natural Context section, with its 240 prosodic units, the coefficient was $k = 0.79$.

In the Telephonic section, however, the overall agreement achieved was lower ($k = 0.66$), but it still fell within the range of values considered substantial (i.e.,

0.61–0.8). With a total of 67 prosodic units, not only were the 20 TSs in the Telephonic section smaller – 3.4 prosodic units per TS on average, while the other two sections taken together had 6.0 units per TS – but they were also less likely to feature a TOP unit than their counterparts in the other sections.

This is because TOP is an information unit characteristic of exchanges that exhibit higher degrees of textual elaboration, that are less interactive and less context dependent. The telephone calls documented in the corpus, however, are for the most part the opposite of that, being quite short and simple in terms of textual structure. Table 6 below summarizes the results discussed so far.

Table 6. Kappa coefficients (*k*), total number of prosodic units (IU), and mean number of TOPs detected by participants (and standard deviation).

Section	***k***	**IU**	**Mean (SD)**
All	0.79	541	52 (2.7)
Media	0.80	234	28 (0.7)
Nat. Context	0.79	240	22 (2.9)
Telephonic	0.66	67	1.8 (0.4)

No annotator found more than 2 TOPs in the Telephonic section and there was one case of full agreement. The kappa statistic is sensitive to the prevalence of a phenomenon, and a small kappa coefficient does not necessarily reflect a low degree of agreement for rare findings (Viera & Garrett 2005). Therefore, the lower coefficient for agreement in the Telephonic section should be interpreted taking that into consideration.

Finally, we computed kappa coefficients in two other ways:

- not replacing any tag with XXX, i.e. leaving the tags as the annotators used them (see column B in Table 6)
- not distinguishing the illocutionary units, i.e. COMs, COBs and CMMs replaced with the ILL tag (see column C in Table 7).

Table 7. The three ways the tags were considered for computing *k*.

Tagged utterance	**A**	**B**	**C**
a tecnologia /=TOP= chegou pra ficar	TOP XXX	TOP COB	TOP ILL
/=COB= ninguém tem dúvida //=COM=	XXX	COM	ILL

The coefficient *k* computed considering the four annotators and the tags under the condition illustrated in column B was $k = 0.72$, which expresses a substantial

agreement. When all the illocutionary units were considered the equal – condition illustrated in column C – the coefficient was a little higher: $k = 0.74$. We take these results as an indication that the general criteria for information structure annotation established by L-AcT are valid and that they can be reliably used for the annotation of speech corpora. Otherwise, the levels of agreement reported in this paragraph would have probably been much lower, considering that the focus of the annotation task was mainly the identification of TOP units.

5. Discussion

In this section, we present a qualitative analysis of the disagreements observed in our reliability test. We will begin with a general discussion of possible motivations for the divergences observed, then we will discuss issues related to confusing TOPs with dialogic units, and finally we will discuss divergences that have a semantic nature.

5.1 Qualitative analysis: possible motivations for the disagreements

In this paper, we have reported the results of an agreement test that supports the validity of the definition of TOP proposed by L-AcT. As discussed earlier, the test indicates substantial agreement among the four participants, who applied the L-AcT definition of TOP to spontaneous speech data.

We examined each disagreement individually, finding 11 instances of TOPs that were mistakenly identified as some other type of information unit and 25 instances of some other information unit being mistakenly identified as TOP. We subjected these inconsistencies to further analysis, which showed basically three general kinds of errors:

- Semantic errors (27 cases, i.e. 75.0%), encompassing the cases in which the ability of a given unit to supply a cognitive domain for the interpretation of the illocution was at stake; these show how important it is to consider both the semantic and the prosodic criteria for the identification of TOP, given that other functions may be conveyed by units exhibiting prosodic features that can be easily confused with those that TOP exhibits (for details about the semantic characteristics of TOP, see section 1.1).
- Syntactically motivated errors (7 cases, i.e. 19.4%): (i) these are cases in which the issue of compositionality between the expression in TOP and that in the illocutionary unit was at stake; the two potentially compositional units may have been interpreted as a SCA-COM sequence, which happened when

their prosodic prominence was not particularly marked; in such cases, it may have been hard to determine whether a nominal expression was functioning as subject or TOP (in some of them both interpretations were reasonable); and an additional case in which the error seems to have been caused by the fact that the expression in TOP was a VP, a syntactic structure that is frequently associated with illocutionary value, particularly when they occur in stanzas, where the illocutions are weakened (Cresti 2009);

- Theoretically motivated errors (2 cases, i.e. 5.6%), which involved cases in which the units – composed of *contudo* 'however' and *afinal* 'after all' – could potentially be functioning as a dialogic unit having a cohesive function (see section 1).

Different factors may have caused these errors, but most of them appear to have been caused by (i) inattention on the part of the annotator as to the semantic aspect of TOP – i.e. the fact TOP must function as a cognitive domain for the illocution –, and (ii) syntactic and prosodic aspects, including speech style, syllable lengthening, f0 movements, short syllabic extension, among other things. In general, monologues, especially those found in media, feature many cases of SCA units whose prosodic and functional characteristics are often a little unclear, making it hard to distinguish whether the expression in SCA is functioning as subject or TOP (in fact, sometimes both interpretations defensible). We also observed a case in which overlapping speech compromised a reliable categorization of the unit.

In the following sections some of these cases will be discussed in more detail.

5.2.1 *Topic or cohesive Discourse Markers?*

We will start out this discussion by looking at the least frequent type of errors, that is, errors that may be theoretically motivated. These were caused by the morphosyntactic makeup of two prosodic units holding expressions that are mostly used with non-referential functions. The expressions in question are *contudo* 'however' and *afinal* 'after all', and they exhibit prosodic features that may look like those that are commonly associated with TOP and can receive two different semantic interpretations: while they can be just a connector establishing a pragmatic, textual relation between two utterances or stanza sub-patterns, they can also supply an identification domain in an anaphoric way. It is not always easy to decide between the two interpretations, and it seems that this is an aspect of the theory that needs to be further investigated.

The expression *contudo* in example (12), whose TS is shown along with two preceding utterances, has anaphoric meaning, which we believe led three annotators

to label it TOP. Our analysis of the prosodic characteristics with which *contudo* is rendered supports this decision.

(12) a olimpíada / trouxe esse legado // houve uma necessidade / houve uma intenção / de fazer um planejamento plurigovernamental // *contudo* /[TOP] muito das ações prometidas / não foram executadas nos prazos adequados //
'the Olympics / brought this legacy // there was a need / there was an intention / to devise a plan involving several governments // *however* / many of the promissed measures / were not taken by the deadlines //'
bmedrp08_2_10-12

As for the other case, shown in example (13) below, despite its prosodic characteristics suggesting it may be a TOP unit – particularly the rising f0 movement with which it ends –, the expression *afinal* has no referential function, hence only one annotator incorrectly labeled it TOP.

(13) *afinal* / quem não gosta / ou não gostaria de ter um carro //
'after all / who doesn't like / or wouldn't like to have a car //'
bmedsc03_3_8

It is worth mentioning that mistakes involving connecting expressions such as these are related to an ongoing discussion about the real status of dialogic units (discourse markers in other frameworks), in particular the unit called *Discourse Connector*, which is a cohesive rather than an interactive dialogic unit (Raso 2014; Raso & Vieira 2016; Raso & Ferrari forthcoming; Cresti & Moneglia 2019; Cresti & Moneglia forthcoming). In fact, the status and the prosodic form of this unit is yet to be fully established, and developments in this respect are likely to reduce the confusion with TOP.

5.2.2 *Syntactically motivated errors*

The analysis of the disagreements suggests explanations for some of the divergences observed that are related to syntactic aspects. The first two prosodic units in example (14) below, taken from a TV news show, are a case in point.

(14) *o ministro da Fazenda* / *Guido Mantega* / culpou o juiz americano / pelo calote argentino / que ele diz que não existiu //
'the Finance minister / Guido Mantega / blamed the American judge / for Argentina's default / which he claims did not happen //'
bmednw03_2_3

All but one annotator used the tag NTP to label the two prosodic units in question, which means that probably interpreted them as a sequence of SCAs. The diverging annotator judged these units to be a sequence of TOPs. By looking at Figure 7 below and listening to the corresponding audio file (ex14.wav), it can be noted that these units (marked (i) and (ii) in Figure 7) cannot be TOPs, given that they have none of the prosodic features of a TOP unit. In fact, they constitute a subject compositional with the VP in the third prosodic unit that happens to be realized by two intonation units.

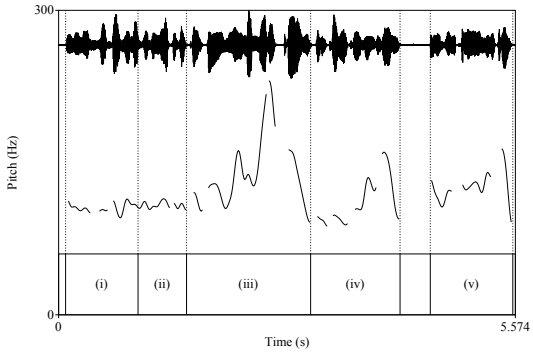

Figure 7. f0 curves of the utterance shown in example (14), with prosodic units numbered (i) to (v) in order to facilitate the identification of relevant portions.

The manner in which these two were realized seem to be correlated with the speech style characteristic of news shows, where presenters tend to overarticulate words and produce prosodic breaks for reasons other than that of signaling the completion of an information unit. To an inattentive annotator, the presence of such breaks may seem to indicate that an information unit has been performed, particularly if the semantic content of the units is construable as an identification domain. We found four cases in which speech style may explain, at least in part, the disagreements observed. These have been counted in the syntactic group of errors, since the breaks induced by speech style always occur at the boundary of a syntactic phrase and given that the possible misinterpretation invariably has to do with the alternative between TOP and subject.

As mentioned earlier, the syntactic errors comprise the disagreements in which the issue of compositionality were at stake. In such cases, an annotator may find it hard to decide whether she is facing a subject in SCA unit or a TOP unit. This problem only arises if the prosodic features of the unit are not completely clear. In fact, 25% of full agreements (i.e. 9 cases) occurred in circumstances

which the expression filling the unit labeled TOP could have been considered syntactically compositional with the following unit. The first prosodic unit in example (15) shows one of such cases where unambiguous prosodic features prevented this type of error from occurring.

(15) *e essa declaração juramentada* /^{TOP} já /^{SCA} restringiu bastante as vendas brasileiras pra Argentina /^{COB} desde dois-mil-e-doze //^{COM}
'and this sworn statement / already / has considerably restricted Brazilian sales to Argentina / since 2012 //'
bmednw03_1_17

Once again, this kind of problem is strictly connected with media style, where SCAs are much more frequent than in other speech styles, especially if compared to informal interactive speech. Another aspect that must be highlighted is that it is not true that the interpretation of a unit as a subject in SCA or as a TOP unit invariably leads to incompatible cognitive interpretations. Cases in which the prosodic form is not clearly marked, different interpretations are possible, without any significant discourse effects.

Finally, there was one case counted in the group of syntactically motivated errors that appears to have been caused by a combination of factors, namely (i) its short syllabic extension, (ii) the presence of adjacent retractions and (iii) the possibility of there being a compositional relationship with expression in the illocutionary unit (hence its being included in the syntactic group of errors). The unit in question was composed simply by the first-person singular subjective pronoun *eu* and one annotator labeled it NTP, while the other three labeled it TOP. It is shown in example (16) below.

(16) agora traz um grande investimento / eu /^{TOP} nũ [/1] nũ [/1] nũ estou aqui mais / pra falar do investimento / (...)
'now it brings a large investment / I / not [/1] not [/1] am not here any more / to talk about the investment / (…)'
bnatpd05_15

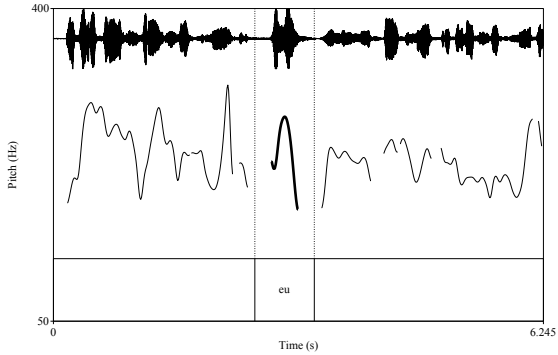

Figure 8. f0 curves of the utterance in example (16) highlighting the portion corresponding to the word "eu".

As the figure and the audio file shows, the unit does resemble a TOP unit, both auditorily and visually. However, not only is it immediately followed by two retractions but is potentially compositional with what comes after it, hence, we believe, the mislabeling.

5.2.3 *Semantic Errors*

Moving on to the semantic errors, we observed 19 cases of expressions that, although incapable of establishing a cognitive domain, were labeled TOP by at least one of the annotators. Our analysis indicated that the types and number of units that were mistakenly identified as TOP were as follows:

- 1 EMP unit (i.e., a retraction)
- 2 PAR units
- 7 SCA units (4 of which being scanned TOPs, 2 COMs, and 1 COB)
- 9 COB units

The EMP (empty) unit, that is, a unit that is retracted and thus has no informational value, was judged to be TOP most probably because the transcription of its rephrasing in one of the following units contained a mistake that had been missed during the revision. Example (17) shows the transcription as sent to the annotators, but the penultimate unit should actually have been transcribed "*embebição e sinérese*" rather than "*em bebição em sinérise*".

(17) (…) a sinérese / &he / &he / &he / &he / em bebição em sinérese / respectivamente //
'the syneresis / &he / &he / &he / &he / imbebition and syneresis / respectively //' [translation corresponds to what was actually pronounced]
bnatte08_113

The cases of PAR units being mislabeled TOP – example (11) above shows one of them – are related to what was discussed at the end of Section 3.1 regarding the prosodic units occurring after TOPs and showing a clear prosodic contrast with its neighboring units. Since the annotators had been advised to watch for such cases in the protocol, we attribute the mislabeling to prosody and partly to inattention. However, this is another case that suggests a need for theoretical refinement, since long PARs – i.e., environments in which it is possible to find units that present characteristics similar to TOP or other specific information units in a level different from the main level of the utterance – have not received enough attention in this respect.

Regarding the SCA units mistakenly labeled TOPs, prosody and inattention appear to have played a part in causing the errors. The four SCA units related to illocutionary units were mislabeled TOP probably due to the presence of slightly rising melodic movements at their final portions, which an attentive inspection considering other prosodic and semantic aspects reveals to be incapable of signaling the function of TOP. As for the SCA units related to TOPs – i.e., SCA units carrying part of the locutive content of TOP units without carrying the prosodic features – inattention appears to be the main cause of errors.

Finally, in those cases where COBs were mislabeled TOPs, most of the errors occurred due aspects related to prosody and disregard on the part of the annotator for the necessary semantic correlate of TOP. As mentioned in section 1, COB is an illocutionary unit that features a prosodic signal of continuity. It happens that this continuity signal sometimes takes the form of a prominent f0 movement coupled with syllable lengthening, and an inattentive annotator can take a signal of continuity of this kind to be the marking of a TOP unit. This is what happened with the COB unit in example (15) above, whose melodic shape is shown in Figure 9 below. The portion of the unit featuring the continuity signal corresponds to the two final syllables of the word *Argentina*.

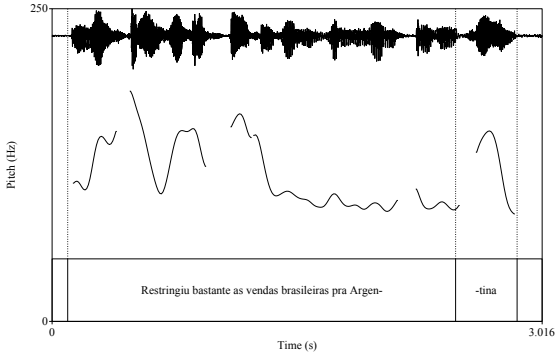

Figure 9. f0 curve of COB unit from the utterance shown in example (15).

As Figure 9 shows, the COB curve from example (11) exhibits a pronounced movement at its end, which we assume led one annotator to label it TOP rather than COB. From the semantic point of view, the expression in COB from the example (11) cannot be the domain for the interpretation of the COM unit. In fact, both the COM and COB units in question must be interpreted according to the domain identified by *this sworn statement*, which is in the prosodic unit that actually functions as TOP.

5.3 *Final words*

We think that our results were much more effective, if compared to other results which aimed to test the notion of topic as semantic aboutness and that do not take in consideration the pragmatic relation between topic and illocution. This is especially relevant if we consider that our data are all extracted from natural spoken texts.

It would be interesting to conduct in the future an inter-annotator agreement test similar to the one reported using a larger sample. A larger sample would be particularly valuable for exchanges with higher degrees of interactivity, given that high interactivity disfavors the occurrence of TOP units, as we have seen with the telephonic sample. In addition, doing the same kind of test but considering other information units, both textual and dialogic, could shed light on other aspects of L-AcT that still needs refinements. The disagreements that have been discussed in the sections above may be indicative of theoretical aspects of L-AcT that may require further investigation both from the prosodic and functional standpoints.

Acknowledgments

This study was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* - Brasil (CAPES) and by the *Fundação de Amparo à Pesquisa de Minas Gerais* (FAPEMIG).

References

Beck, K. 2012. Tübinger Baumbank des Deutschen/Zeitungskorpus (TüBa-D/Z). *Seminar for Linguistics*. Eberhard Karls University of Tübingen, Tübingen. https://uni-tuebingen.de/fakultaeten/philosophische-fakultaet/fachbereiche/neuphilologie/seminar-fuer-sprachwissenschaft/arbeitsbereiche/allg-sprachwissenschaft-computerlinguistik/ressourcen/corpora/tueba-dz.html (accessed April 30, 2020).

Boersma, P. & Weenink, D. 2019. *Praat: doing phonetics by computer: computer program*. Version 5.4.21 Amsterdam: University of Amsterdam. http:// www.praat.org/ (accessed February 2020).

Bossaglia, G. & Ferrari, L. A. 2019. The C-ORAL-BRASIL project: varied resources for the study of spoken Brazilian Portuguese. *The Journal Of Speech Sciences* 7(2), 65-77.

Cavalcante, F.A. 2015. *The topic unit in spontaneous American English: a corpus-based study*. M.A. thesis, Faculdade de Letras, Universidade Federal de Minas Gerais.

Cavalcante, F.A. 2020. *The information unit of Topic: a crosslinguistic, statistical study based on spontaneous speech corpora*. PhD diss., Faculdade de Letras, Unversidade Federal de minas Gerais.

Cavalcante, F. A. & Ramos, A., 2016. The American English spontaneous speech minicorpus. Architecture and comparability. *CHIMERA: Romance Corpora and Linguistic Studies 3*(2), p.99-124.

Cavalcante, F. A., Raso, T., & Ramos, A. 2018. American English Informationally Tagged Minicorpus. http://www.c-oral-brasil.org.

Chafe, W.L. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C.N. Li (ed). *Subject and topic*. New York: Academic Press.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20(1), 37-46.

Cresti, E. 2000. *Corpus di italiano parlato*. Firenze: Accademia della Crusca.

Cresti, E. 2009. La Stanza: un'unità di costruzione testuale del parlato. X Congresso SILFI, 2009, Basileia. *Atti del X Congresso SILFI: Sintassi storica e sincronica dell'italiano. Subordinazione, coordinazione e giustapposizione*. Basileia, 1-25.

Cresti, E. 2011. The definition of focus in Language into Act Theory (LAcT). In: Mello, H., Panunzi, A., Raso, T. (org.). *Pragmatics and Prosody: Illocution Modality, Attitude, Information Patterning and Speech Annotation*. Firenze: Firenze University Press.

Cresti, E. 2014. Syntactic properties of spontaneous speech in the Language into Act Theory: data on Italian complements and relative clauses. In T. Raso, H.R. MELLO (eds), *Spoken corpora and linguistic studies*. Amsterdam: John Benjamins, p. 365-410.

Cresti, E. 2018. The illocution-prosody relationship and the Information Pattern in spontaneous speech according to the Language into Act Theory (L-AcT). *Linguistik Online*, *88*(1).

Cresti, E. & Moneglia, M. (eds) 2005. C-ORAL-ROM. *Integrated reference corpora for spoken Romance languages*. Amsterdam: John Benjamins.

Cresti, E. & Moneglia, M. 2019. *The Discourse Connector according to the Language into Act Theory: data from IPIC Italian*. In: Bidese, E., Casalicchio, J. & Moroni, M. (eds.): *La linguistica vista dalle Alpi. Teoria, lessicografia e multilinguismo / Linguistic views from the Alps. Language Theory, Lexicography and Multilingualism*. Frankfurt am Main, Peter Lang, 99-126.

Cresti, E. & Moneglia, M. forthcoming. Il Connettore discorsivo secondo la Teoria sulla lingua in atto. In A. De Meo & F. Dovetto (eds.), Atti del Congresso GSCP "La comunicazione parlata" Università degli Studi di Napoli "L'Orientale" - Università degli Studi di Napoli Federico II (Napoli, 12-14 dicembre 2018), Napoli: Aracne.

Cook, P. & Bildhauer, F. 2013. Identifying "aboutness topic": two annotation experiments. *Dialogue and Discourse* 4(3), 118-141.

Davies, M., Fleiss, J. L. 1982. Measuring agreement for multinomial data. *Biometrics* 38(4), 1047-1051.

Du Bois, J. W., Chafe, W. L.; Meyer, C.; Thompson, S. A.; Englebretson, R.; Martey, N. 2000-2005. *Santa Barbara corpus of spoken American English*, Parts 1-4. Philadelphia: Linguistic Data Consortium.

Firenzuoli, V. 2003. *Le forme intonative di valore illocutivo dell'italiano parlato: analisi sperimentale di un corpus di parlato spontaneo (LABLITA)*. PhD diss. Università di Firenze.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76(5).

Frosali, F. 2008. L'unità di informazione di ausilio dialogico: Valori percentuali, caratteri intonativi, lessicali e morfo-sintattici in un corpus di italiano parlato (C-ORAL-ROM). In E. Cresti (ed.), *Prospettive nello studio del lessico italiano*. Florence: Firenze University Press, 417-424.

Gamer, M., Lemon, J., Singh, I.F.P. 2019. *irr: Various Coefficients of Interrater Reliability and Agreement*. R package, version 0.84.1. https://CRAN.R-project.org/package=irr (accessed March 10, 2020).

Gobbo, O. 2019. *Marcadores discursivos em uma perspectiva informacioanl: análise prosódica e estatística*. M.A. thesis, Faculdade de Letras, Universidade Federal de Minas Gerais.

Grønnum, N. 2006. DanPASS - A Danish Phonetically Annotated Spontaneous Speech corpus. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Tapias, D. *Proceedings of the 5th International Conference on Language Resources and Evaluation, Genova 24-26 May 2006.* European Language Resources Association, Genova.

Krifka, M. 2008. Basic notions of information structure. *Acta Linguistica Hungarica*, 55(3-4), 243-276.

Krippendorff, K. H. 2004a. *Content analysis: an introduction to its methodology*. Thousand Oaks: SAGE Publications.

Krippendorff, K. 2004b. Reliability in Content Analysis. *Human Communication Research* 30(3), p. 411-433.

Kupietz, M., Keibel, H. 2009. The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In Minegishi, M., Kawaguchi, Y. (eds.), *Working Papers in Corpus-based Linguistics and Language Education*, *3*. Tokyo: Tokyo University of Foreign Studies (TUFS), 53-59.

Lambrecht, K. 1994. *Information structure and sentence form: topic, focus and the mental representations of discourse referents*. Cambridge: Cambridge University Press.

Landis, R.J. & Koch, G.G. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*: 363–374.

Li, C.N., Thompson, S.A. 1976. Subject and topic: a new typology of language. In C.N. Charles (ed). *Subject and topic*. New York: Academic Press.

Lüdeling, A., Ritz, J., Stede, M., Amir, Z. 2016. Corpus linguistics and information structure research. In C. Féry, S. Ishihara (eds). *The Oxford Handbook of Information Structure*. Oxford: Oxford University Press, 599-620.

Maia Rocha, B. & Raso, T. 2011. A unidade informacional de introdutor locutivo no português do Brasil: uma primeira descrição baseada em corpus. *Domínios de Linguagem*.

Mello; H.R. 2014. Methodological issues for spontaneous speech corpora compilation: the case of the C-ORAL-BRASIL. In T. Raso, H.R. MELLO (eds), *Spoken corpora and linguistic studies*. Amsterdam: John Benjamins, 29-68.

Mittmann, M. M. 2012. *O C-ORAL-BRASIL e o estudo da fala informal: um novo olhar sobre o tópico no português brasileiro*. PhD diss., Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte.

Moneglia, M., Raso, T. 2014. Notes on Language into Act Theory (L-AcT). In T. Raso, H.R. MELLO (eds), *Spoken corpora and linguistic studies*. Amsterdam: John Benjamins, 469-495.

Paggio, P. 2006. Annotating information structure in a corpus of spoken Danish. *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (LREC'06), 1606-1609. http://www.lrec-conf.org/proceedings/lrec2006/pdf/639_pdf.pdf (accessed April 2020).

Raso, T. 2012. O corpus C-ORAL-BRAISL. In In T. Raso, H.R. MELLO (eds), *C-ORAL-BRASIL I: corpus de referência de português brasileiro falado informal*. Belo Horizonte: UFMG, 55-90.

Raso, T. 2014. Prosodic constraints for discourse markers. In T. Raso, H.R. MELLO (eds), *Spoken corpora and linguistic studies*. Amsterdam: John Benjamins, 411-467.

Raso, T. & Ferrari, L.A. forthcoming. Uso dei Segnali Discorsivi in corpora di parlato spontaneo italiano e brasiliano. In: Ferroni, R., Birello, M. (eds.) 2020. *La competenza discorsiva a lezione di lingua straniera.* Roma: Aracne.

Raso, T., Mello, H.R. (eds) 2012. *C-ORAL-BRASIL I: corpus de referência de português brasileiro falado informal*. Belo Horizonte: UFMG.

Raso, T., Mello, H.R., Ferrari, L.A. forthcoming. *C-ORAL-BRASIL II: corpora of Brazilian Portuguese speech in formal, media, and telephonic interactions*.

Raso, T., Cavalcante, F., Mittmann, M. 2017. Prosodic forms of the Topic information unit in a cross-linguistic perspective: a first survey. Proceedings of the *SLI-GSCP International Conference*, 13-15 June, 2016. A. de Meo & F. M. Dovetto (eds), Rome: Aracne editrice, 473-498.

Raso, T. & Rocha, B. 2017. Illocution and attitude: on the complex interaction between prosody and pragmatic parameters. *JOSS Journal Of Speech Science*, 5, 5-27.

Raso, T., Vieira, M. 2016. A description of Dialogic Units/Discourse Markers in spontaneous speech corpora based on phonetic parameters. *Chimera: Romance Corpora and Linguistic Studies*, 3, 221-249.

R Development Core Team. 2019. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org (accessed February 2020).

Ritz, J., Dipper, S., Michael, G. 2008. Annotation of information strucute: an evaluation across different types of texts. *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC'08). http://www.lrec-conf.org/proceedings/lrec2008/ (accessed April 2020).

Skopeteas, S. Fiedler, I., Hellmuth, S., Schwarz, A., Stoel, R., Fanselow, G., Féry, C., Krifka, M. 2006. Questionnaire on Information Structure (QUIS). In *Working Papers of the SFB632*, Interdisciplinary Studies on Information Structure (ISIS) 4. Potsdam: Universitätsverlag Potsdam.

Stede, M. 2004. The Potsdam Commentary Corpus. Proceedings of the ACL 2004 Workshop on Discourse Annotation, 96-102.

Tucci, Ida. 2010. "Obiter dictum". La funzione informativa delle unità parentetiche. In Atti del Convegno Internazionale GSCP *La comunicazione parlata*, 2009. M. Pettorino; A. Giannini; F. Dovetto (orgs.), Napoli: Università degli Studi di Napoli l'Orientale, p. 635–654

Viera, A.J. & Garrett, J.M. 2005. Understanding interobserver agreement: the kappa statistic. *Fam med* 37(5), 360-363.