Automatic detection of other-repetition occurrences: application to French conversational Speech

Brigitte Bigi, Roxane Bertrand, Mathilde Guardiola

Laboratoire Parole et Langage, CNRS, Aix-Marseille Université,
5, avenue Pasteur, 13100 Aix-en-Provence, France
{brigitte.bigi,roxane.bertrand,mathilde.guardiola}@lpl-aix.fr

Abstract

This paper investigates the discursive phenomenon called other-repetitions (OR), particularly in the context of spontaneous French dialogues. It focuses on their automatic detection and characterization. A method is proposed to retrieve automatically OR: this detection is based on rules that are applied on the lexical material only. This automatic detection process has been used to label other-repetitions on 8 dialogues of CID - Corpus of Interactional Data. Evaluations performed on one speaker are good with a F1-measure of 0.85. Retrieved OR occurrences are then statistically described: number of words, distance, etc.

Keywords: annotation; automatic; other-repetition

1. Introduction

This paper investigates the discursive phenomenon called other-repetitions (OR). Other-repetition is a device involving the reproduction by a speaker of what another speaker has just said. Other-repetition has been identified as an important mechanism in face-to-face conversation through their discursive or communicative functions (Johnstone, 1987; Norrick, 1987; Tannen, 1989; Perrin et al., 2003). Among their various functions in discourse, repetition serves the purpose of facilitating comprehension by providing less complicated discourse, while also establishing connection with earlier discourse (cohesion), or yet also functions as a device for getting or keeping the floor (Norrick, 1987).

There are a number of studies which investigate the OR's functions, just a few are related to their form. This paper proposes to extend and clarify the lexical description of other-repetitions. We focus on a lexical study for the automatic detection and their characterization. An automatic method is proposed to retrieve other-repetition occurrences. This automatic detection (particularly in a spontaneous dialogue) is a challenge as, to our knowledge, it does not already exist such a system. An automatic detection system of self-repetitions in a Human-Machine dialogue is presented in (Bear et al., 1992). It aims at highlighting repetitions as for example "show me *flights* daily *flights* to Boston", with a method based on a two-stages process. Firstly, a set of candidates are proposed by using a pattern matching search. Secondly, information from syntax, semantic and acoustic levels are used to filter these candidates and so to find those relevant. From the proposition in (Bear et al., 1992), we kept the idea of a two-steps algorithm to find other-repetitions between two speakers in a conversation. Then, the first step consists in finding a set of candidates: words, or word sequences of the source speaker matching with words pronounced by the echoing-speaker. The second step consists in establishing rules to accept or reject these candidates according to identification criteria of the OR. A key-point is that the proposed automatic detection is based on observable cues which can be useful for OR's identification from the transcription. Furthermore, this tool was used to propose a lexical characterization of OR: various statistics are estimated on the detected OR. Indeed, the detection process has been used to label CID - Corpus of Interactional Data (Bertrand et al., 2008). This corpus is an audio-visual recording of 8 hours of French conversational dialogues (1 hour of recording per session). Each audio signal (one speaker) is automatically segmented in IPUs - Inter-Pausal Units. IPUs are blocks of speech bounded by silent pauses over 200 ms, and aligned on the speech signal. For each of the speakers an orthographic transliteration is provided which is used in this work. The transcription process was done following specific conventions derived from GARS (Blanche-Benveniste and Jeanjean, 1987). Each dialogue involves two participants of the same gender. One of the following two topics of conversation was suggested to the participants: conflicts in their professional environment or unusual situations in which they may have been. These instructions were not exhaustive and participants often spoke very freely about various topics, in a conversational speaking style.

The proposed method to automatically detect other-repetitions is described in the next section. The evaluation of such system is then proposed. Finally, a description of the whole set of the collected repetitions on CID is proposed: the formal characteristics of ORs are investigated. In previous studies, CID was richly annotated (see (Blache et al., 2010)) and some annotations are distributed for research purposes[1]. Then, the OR occurrences will also be distributed.

2. Automatic detection: Method

2.1. Preliminary study

Tannen (1989) described other-repetition in conversation, distinguishing exact repetition and repetition with variation (including various variation such as prosodic variation or reformulation). We here exclude reformulation and we concentrate on verbal repetition (with the same words). A broader, more formal repetition was proposed by (Chiungchih, 2010) as exact, reduced, modified or expanded repetition.

[1] http://www.sldr.fr/sldr000720

Prior to the automatic method development an expert has manually annotated the whole OR occurrences on one dialogue to characterize the various types of observed other-repetition in a spontaneous dialogue. It allowed to fix some lexical cues. We identified 3 main properties. Firstly, we observed word variations as singular/plural, a pronoun variation or a tense change. Another type of frequent observed variation was words inserted in the repeated sequence or words not repeated in the same order (for example: *the green horse / the horse is green*). Finally, another characteristic of other-repetitions concerns the distance between the repeated words and their source. By opposition to distant-repetitions, local repetitions are usually expressed as a simple echo of the immediately prior talk (Perrin et al., 2003). However, this manual annotation showed that an other-repetition can appear much later in the dialogue.

2.2. Finding a set of candidates

The automatic detection focus on word repetitions, which can be an exact repetition (named strict echo) or a repetition with variation (named non-strict echo). Repetitions with variations, which are the most problematic, implies solving different problems mentioned in the previous section.
Firstly, it is preferable to get sources instead of echos, such as the example:

CM et *il contrôlait pas*
AB *il* a *pas contrôlé*

CM and he was not controlling
AB he has not controlled

Word insertions are very frequent. Detecting the source allows to get the entire set of words of the sequence: in the example, detecting the echo implies to miss the word "il".
Secondly, variations such as singular/plural of the same word, pronoun variation or change of tense was solved by the use of lemmas. Here is an example[2] of word variations:

EB c'était quand je bossais *en Belgique*
SR ah oui c'est vrai tu as bossé *en Belgique*

EB it was when I was working in Belgium
SR ah yes that's right you have been working in Belgium

This example was lemmatized as:

EB ce être quand *il bosser en Belgique*
SR ah oui ce être vrai *il bosser en Belgique*

Consequently, the automatic detection based on lemmas produced the sequence of 4 lemmas *il bosser en Belgique*. In the following, the use of the term "word" will refer to the lemma of the word.
Another problem was to define the time length in order to find repeated items in the dialogue. We propose to fix this length on the basis of the IPU segmentation. The automatic other-repetition detection consists in matching lemmas of the speaker in a given IPU with lemmas of the other-speaker in the same time-localization IPU and then in the N following IPUs. Then the time length to find repeated items is variable as IPUs have a different duration.
These processing provides the entire set of text segments which are repeated. Obviously, this set must be filtered. Figure 1 illustrates an example of automatic detection. The processing of the algorithm produces all the boxes drawn in the source (those below). The second processing step aims to select only ones which are relevant (square boxes) and reject the others (round boxes).

2.3. Selecting candidates

The aim is to keep all of the real sources of other-repetitions from the set of repeated items while removing a maximum of false ones (simple matching items or other types of repetitions). A set of rules was defined to examine each candidate. The proposed rules are the result of discussions with experts held prior to the development of the automatic tool. Proposed rules deal with the number of words, the word-frequencies and distinguish if the repetition is strict or not. The following rules are proposed:

Rule 1 A source is accepted if it contains one or more relevant word. Relevance depends on the speaker producing the echo;

Rule 2 A source which contains at least K lemmas is accepted if the repetition is identical.

Rule number 1 needed to fix a clear definition of the relevance of a word. A fixed list of stop-words could be used, where a word is relevant if it does not occurs in this list. However, in a dialogue corpus with spontaneous speech and open topics, we suggest that a better choice is to fix this list from words observed in the dialogue. Because, a word can be relevant in a dialogue and not in an other, or not in the language in general. Moreover, we observed that both speakers of a dialogue are using their own vocabulary and relevant words are different from each other. Then, if the dialogue contains enough data, a list of relevant words can be estimated independently for each speaker.
Let $N_l(w)$, the number of occurrences of the word w of the speaker l, and $|V_l|$, the vocabulary size (number of different words) of the speaker l. Let then $P_l(w)$, the probability of the word w of the speaker l, defined by:

$$P_l(w) = \frac{N_l(w)}{\sum_i^{|V_l|} N_l(w_i)}$$

A word w is relevant for the speaker l if its probability is less than a threshold. It depends on the speaker vocabulary:

$$P_l(w) \leq \frac{1}{\alpha \times |V_l|}$$

The α value could be empirically estimated, depending on the corpus.
For example, applying the rules on the example described in Figure 1 will select only the candidate *"c' était un bar"* by the use of the rule 1 (*"bar"* is relevant) or the rule 2 (echo strict more or equal than 3 words). The two others candidates are rejected: too shorts and without a relevant word.

[2] We note the speakers in a bold font. Words/Lemmas which are repeated are written in an italic font.

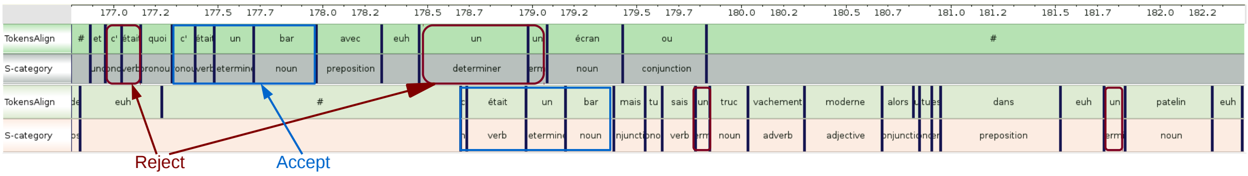

Figure 1: *Other-Repetition detection example*

2.4. Getting the repetition

The proposed algorithm is focusing on the source detection. If there are more than one possible repetition, or if multiple repeated segments are found, the algorithm to get a repetition for each source is following this rule: the longest (from left to right) then, the nearest.

For example, the lemmatized sequence of the 4 tokens *c'était un bar*, which is described in the Figure 1, is found as follow:

- **ce** is found at index 0 in the echoing-speaker;
- **être** is found at index 1 and 14;
- **un** is found at index 2, 7, and 17;
- **bar** is found at index 3.

At the first iteration, the algorithm returns the sequence 0,1,2,3 (the longest) that covers all the lemmas of the source. Consequently, the algoritm stops and return the repetition made of the sequence "ce être un bar" (strict echo).

Another example is given in the following (AB is the source, and the content of the sequence of the 5 next IPUs of the echoing speaker CM):

AB *le petit feu de* artifice ouais ce être le tout petit truc là

CM *le feu* # ah *le petit* machin *de* ouais ouais ouais ouais ouais d'accord ouais + ouais ouais ouais ouais + hum hum ouais hum hum @@ # ouais # ouais ouais ouais ouais + hum # ouais oui oui ce être pas *le le* ouais ah ouais ouais @@

AB the little fireworks yeah it was that little thing

CM the fire [sil] ah the little thing of yeah yeah yeah yeah yeah okay yeah [sp] yeah yeah yeah yeah [sp] hum hum yeah hum hum [laugh] [sil] yeah [sil] yeah yeah yeah yeah [sp] hum [sil] yeah yes yes it wasn't the the yeah ah yeah yeah [laugh]

- **le** is found at index 0, 4, 44 and 45 in the echoing-speaker;
- **petit** is found at index 5;
- **feu** is found at index 1;
- **de** is found at index 7 and 13.

The resulting repetitions sequences are:

- **le petit** at index 4 to 5;
- **feu** at index 1;
- **de** at index 7.

This solution ensure to get all the lemmas of the source in the repetition and expect to get the most appropriate segments if the repetition is not strict.

3. Results

3.1. Implementing in a tool

The system proposed in this paper is implemented in SPPAS (Bigi and Hirst, 2012), a tool distributed under the terms of the GNU Public License[3]. The OR-detection system is proposed in the form of a Python program, freely available in the *bin* directory.

This program was used to detect other-repetitions of the 8 dialogues of CID. The input files contain one tier with the orthographic transcription of the IPUs of each speaker involved in the dialogue. Orthographic Transcription is then time-aligned at the token level with the help of SPPAS (Bigi, 2012). Moreover, the POS-Tagger MarsaTag[4] is also applied on the data (as can be shown in Figure 1). The other-repetition occurrences (sources and echos) on the whole corpus will be publicly available in the form of TextGrid files on the SLDR[5].

3.2. CID: lexical description

The vocabulary, the number of occurrences of lemmas and the number of hapax[6] used by each speaker are presented in table 1. Each horizontal line of the table is a dialogue separator, and speakers are mentioned with their initials. For each of them, we can observe a high rate of hapax: from 46% to 55%. The complete vocabulary is about 48K lemmas, but only 111 lemmas are pronounced (at least one time) by all speakers.

By applying the definition of relevance we propose on the AB-CM dialogue of CID, a lemma of the speaker AB is not relevant if it occurs more than 15 times. It represents only 60 lemmas of her vocabulary, but 71.32% of her word occurrences in the dialogue. For CM, a lemma is not relevant if it occurs more than 19 times, which corresponds only to 70 different lemmas but 74.78% of her occurrences in the dialogue. The intersection between the irrelevant lemmas of the two speakers is 55 lemmas, as for example: *ah, aller, avec, avoir, truc, ouais, oui, mais, devoir, pouvoir, ça* ("ah, to go, with, to have, stuff, yeah, yes, but, to have to, can, this). Otherwise, the lemma *petit* ("little") occurs 21 times for AB which make it irrelevant. This lemma occurs only 8 times for CM which make it relevant for her. Otherwise, the lemma *voilà* ("that's it!") is relevant for speaker AB by occurring 8 times, and irrelevant for speaker CM with 27 occurrences. It is the same for the lemma *vachement* ("really") with 2 occurrences for speaker AB and 20 for CM.

[3]See: http://www.gnu.org/licenses/gpl-3.0.en.html for details
[4]http://sldr.org/sldr000841
[5]http://sldr.org/sldr000720
[6]hapax: terms for which the number of occurrences is 1

Table 1: CID: Lemmas-vocabulary description

Spk.	Vocab.	Occ.	Hapax
AB	874	6642	447
CM	783	7878	360
AC	788	6890	369
MB	1210	9560	650
AG	847	7748	433
YM	852	8430	453
AP	1056	8853	578
LJ	1052	9024	580
BX	800	6001	393
MG	952	8346	481
EB	893	6805	467
SR	650	6065	323
IM	980	7633	502
ML	790	6717	375
LL	422	3501	196
NH	831	6789	429

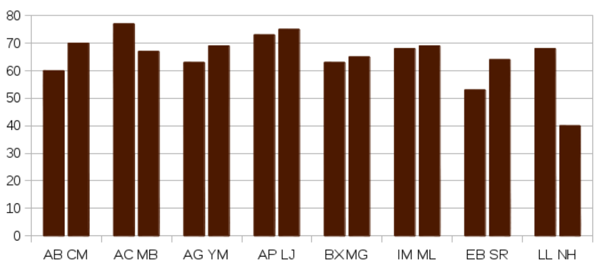

Figure 2: Number of stop words

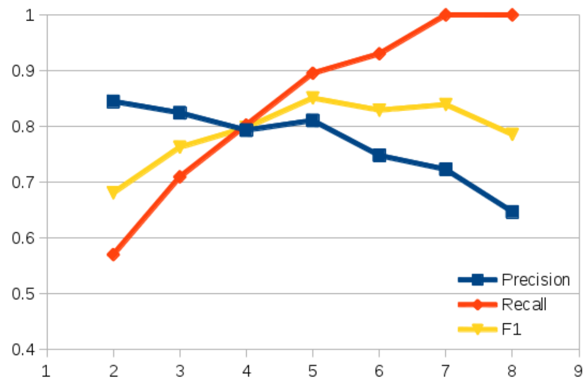

Figure 3: Evaluation with α=0.5

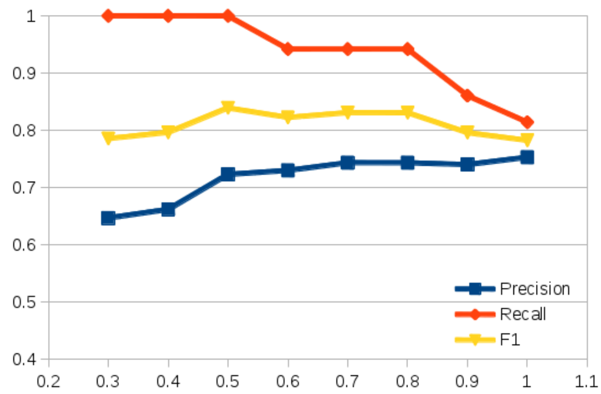

Figure 4: Evaluation with N=7

Figure 2 indicates the number of stop words selected for each speaker.

3.3. OR's detection: Evaluation

One speaker (ie 1 hour speech) was manually annotated by selecting all candidates proposed by the first step of the system, before applying rules. The value used to generate candidates was $N = 9$ to ensure to get the larger set of candidates as possible. The recall, precision and F1-measure was estimated by comparing the system output selection with this manual selection.

Figure 3 shows the results by fixing α=0.5 and by ranging the N value from 2 to 9. The best F1 value is 0.85, which represents a pretty good score given the fact that we offer the first automatic system to detect OR in a dialogue. It is obtained with $N = 5$, and the best recall value with $N = 7$. This confirms that a significant number of other-repetitions occurs much later in the dialogue. Figure 4 shows the results by fixing N=7 and by ranging the α value from 0.3 to 0.1. The best F1 value is observed with α=0.5 as expected.

We also verified if the use of lemma is appropriate, by running the system with words. With $N = 7$, we get recall=0.779 and precision=0.651; and with $N = 5$, we get recall=0.698 and precision=0.706. These results confirm that the use of lemma is suitable.

The last evaluations aim to validate our proposal to create a list of stop words for each speaker. We downloaded a stop words list from the web[7] made of 126 words, and executed our system by using this list for each speaker instead of our proposal. By using $N = 7$, we get recall=0.977 and precision=0.198; and with $N = 5$, we get recall=0.872 and precision=0.223. These results are significantly lower than those presented in Figure 3. We also constructed a list of stop words by using the 65[8] most frequent words in all dialogues. With $N = 7$, we get recall=1 and precision=0.566; and with $N = 5$, we get recall=0.895 and precision=0.636. These results are better than using a general stop list but the precision is significantly lower than creating a specific list for each speaker with the proposed method, as results in Figure 4.

3.4. Examples

The example described below and in Figure 5 is an illustration of the system output.

[7] http://www.ranks.nl/stopwords/french.html

[8] In our proposal, with $N = 5$ and $\alpha = 0.5$, the average number of stop words is 65.

AB ils voulaient qu'on fasse *un feu d'artifice* en fait dans un voy- *un foyer un foyer* catho *un foyer de bonnes soeurs*

CM *un feu d'artifice*

AB ah ouais

CM *dans un foyer de bonnes soeurs*

AB they wanted we made fireworks actually in a Catholic boarding school a nuns boarding school

CM *fireworks*

AB ah yeah

CM *in a nuns boarding school*

By considering the speaker AB as the source and CM the echoing speaker, the system outputs the following sources and repetitions::

- S18, corresponding to AB: *un feu d'artifice*
- S19, corresponding to AB: *un foyer un foyer*
- S20, corresponding to AB: *dans un foyer de bonnes soeurs*
- R18, corresponding to CM: *un feu d'artifice*
- R19, corresponding to CM: *un foyer*
- R20, corresponding to CM: *dans un foyer de bonnes soeurs*

In the next example, the rule 2 is suitable since it enables to achieve the detection of a sequence of 8 irrelevant lemmas:

IM *jusqu'à ce qu' y en ait une qui* réagisse

ML *jusqu'à ce qu' y en ait une qui* veuille bien mais comme euh ils sont quand même cent cinquante enfants

IM until one of them reacts

ML until one of them agrees but as if they are 150 infants

The last example combines several phenomena (irrelevant lemma, inserted lemmas in the echo - *je sais pas* -, displacement of the lemma - *pour* -)

CM *ah ils vous ont pris pour* des rustres peut-être alors hein

AB *ah* je sais pas *pour* quoi *ils nous ont pris* mais nous on s'est dit mais qu'est-ce qu'on est venu foutre là-dedans et

CM ah they think you boorish then well eh

AB I don't know for what they think we are but we think but what we came to do in it and

4. Statistics about other-repetitions

This section presents a set of statistics about the extracted OR occurences, detected with $N = 5$ and $\alpha = 0.5$. As shown in Figure 6, a set of 1711 sources is proposed, with an average of 2.7 words per occurrence (SD=1.65), see Figure 7. The minimum number of words is 1, the maximum is 15. In both figures, speakers are grouped by dialogs. The distance between the source and the repetition is presented in Figure 8.

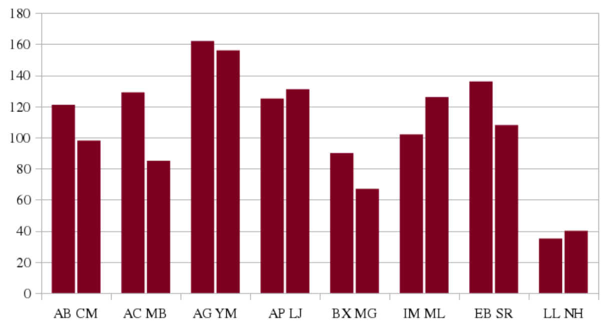

Figure 6: Number of echos per speaker

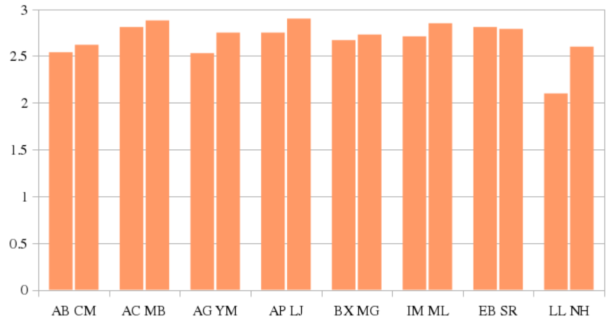

Figure 7: Average number of words per echo

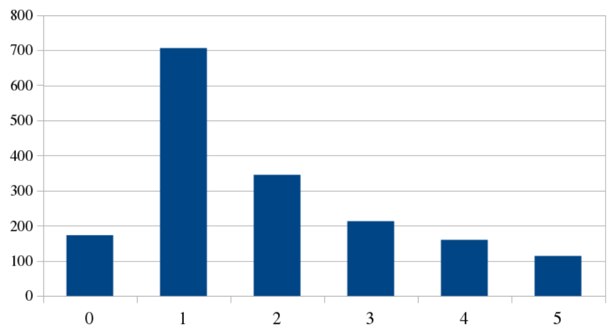

Figure 8: Average distance between the source and the echo

Because the POS tagger was applied on the whole data, we can get the category of each token of the OR (see Table 2). It is interesting to notice that nouns and determiners occurs proportionally more often than the other categories.

In the description of the method, we introduced a list of variations we are facing on while detecting OR occurrences. In Figures 9, 10 and 11, four types of echos are referenced:

- strict: the source and the repetition are strictly identical, at the word level;
- variation: the source and the repetition are identical, at the lemma level;
- reduction: the repetition is shorter than the source;
- split: the echo is piecewise.

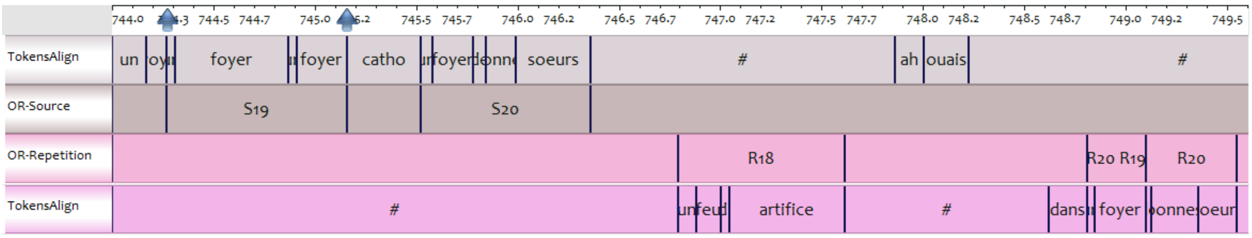

Figure 5: Screenshot of the system output

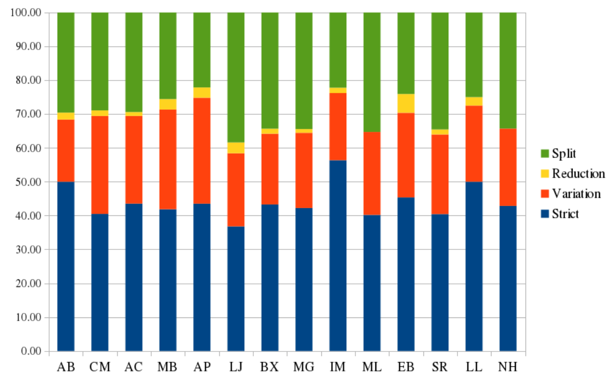

Figure 9: Percentage of each type of echo per speaker

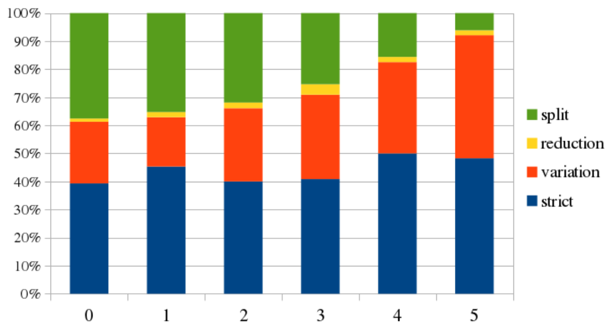

Figure 10: Percentage of each type of echo per distance

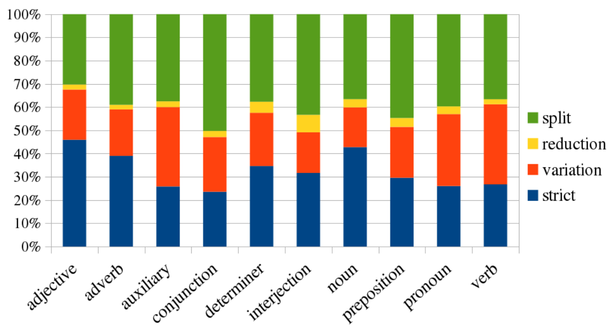

Figure 11: Percentage of each type of echo per category

Table 2: Categories of the sources

Category	# in CID	# in OR	%
adjective	4480	185	4.13
adverb	12338	308	2.50
auxiliary	2964	122	4.12
conjunction	8989	191	2.12
determiner	10058	591	5.88
interjection	8118	120	1.48
noun	13149	798	6.07
preposition	9022	340	3.77
pronoun	26159	1057	4.04
verb	20374	885	4.34
Total	115651	4597	3.97

5. Conclusion

Work related to other-repetitions mainly concerns their functions, but there is a weaknesses on their formal definition. This study on automatic detection of other-repetition described an original method to determine which formal criteria are best, as well as presenting and evaluating the tool we created for this detection and we tested on a French conversational corpus.

Current studies focus on the analysis of the collected OR occurrences. Rich annotations of CID lead us to highlight specific patterns of such OR at syntactic, discursive and prosodic levels. Thanks to a formal analysis of these OR, we will better characterize them.

6. References

J. Bear, J. Dowding, and E. Shriberg. 1992. Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *30th annual meeting on Association for Computational Linguistics*, pages 56–63, Newark, Delaware.

R. Bertrand, P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde, and S. Rauzy. 2008. Le CID - Corpus of Interactional Data. *Traitement Automatique des Langues*, 49(3):105–134.

B. Bigi and D. Hirst. 2012. SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody. In ISBN 978-7-5608-4869-3 Tongji University Press, editor, *Proc. of Speech Prosody*, pages 19–22, Shanghai (China).

B. Bigi. 2012. SPPAS: a tool for the phonetic segmentation of speech. In *Language Resource and Evaluation Conference*, pages 1748–1755, ISBN 978–2–9517408–7–7, Istanbul (Turkey).

P. Blache, R. Bertrand, B. Bigi, E. Bruno, E. Cela, R. Espesser, G. Ferré, M. Guardiola, D. Hirst, E.-P. Magro, J.-C. Martin, C. Meunier, M.-A. Morel,

E. Murisasco, I. Nesterenko, P. Nocera, B. Pallaud, L. Prévot, B. Priego-Valverde, J. Seinturier, N. Tan, M. Tellier, and S. Rauzy. 2010. Multimodal annotation of conversational data. In *The Fourth Linguistic Annotation Workshop*, pages 186–191, Uppsala, Sueden.

C. Blanche-Benveniste and C. Jeanjean. 1987. *Le français parlé*. Transcription et édition, Didier Erudition.

H. Chiung-chih. 2010. Other-repetition in mandarin child language: A discourse pragmatic perspective. *Journal of Pragmatics*, 42(3):825–839.

B. Johnstone. 1987. An introduction. *Text - Interdisciplinary Journal for the Study of Discourse*, 7(3):205–214.

N.R. Norrick. 1987. Functions of repetition in conversation. *Text - Interdisciplinary Journal for the Study of Discourse*, 7(3):245–264.

L. Perrin, D. Deshaies, and C. Paradis. 2003. Pragmatic functions of local diaphonic repetitions in conversation. *Journal of Pragmatics*, 35:1843–1860.

D. Tannen. 1989. *Talking voices: Repetition, dialogue, and imagery in conversational discourse*. Cambridge/New York: Cambridge University Press.