Paraphrastic Reformulations in Spoken Corpora

Iris Eshkol-Taravella, Natalia Grabar

Paraphrastic reformulations in spoken corpora

Iris Eshkol-Taravella[1], Natalia Grabar[2]

(1) CNRS UMR 7270 LLL, Université d'Orléans, 45100 Orléans, France
iris.eshkol@univ-orleans.fr
(2) CNRS UMR 8163 STL, Université Lille 3
59653 Villeneuve d'Ascq, France
natalia.grabar@univ-lille3.fr

Abstract. Our work addresses the automatic detection of paraphrastic reformulation in French spoken corpora. The proposed approach is syntagmatic. It is based on specific markers and the specificities of the spoken language. Manual multi-dimensional annotation performed by two annotators provides fine-grained reference data. An automatic method is proposed in order to decide whether sentences contain or not paraphrastic relations. The obtained results show up to 66.4% precision. Analysis of the manual annotations indicates that few paraphrastic segments show morphological modifications (inflection, derivation or compounding) and that the syntactic equivalence between the segments is seldom respected, as these usually belong to different syntactic categories.

Keywords: Reformulation, Paraphrase, Spoken corpora

1 Introduction

The acquisition of paraphrases is an important research topic in the NLP area, as the paraphrase plays several intra and inter-speaker functions in language:

- it guarantees the natural character and the beauty of language, as it avoids repetitions and redundancies for instance;
- it helps the understanding and communication [1, 2], as it is widely used for the interpretation of religious, philosophical and literary texts;
- in language learning, in indicates the capacity to master the language;
- but it can also prevent the clarity of the communication [3], such as happens in the specialized fields with the technical paraphrases that cannot be easily understood by laymen.

In any case, the speakers must share common knowledge and background to be able to detect and understand the paraphrases, and to appreciate them. For NLP applications, the paraphrase remains a real challenge for the same reasons: it involves a great variety of linguistic and referential mechanisms in both detection and production of paraphrases. Several NLP applications are concerned with the use of paraphrases:

- for information retrieval and extraction, the paraphrases allow accessing more complete and exhaustive information in documents despite the surface dissimilarity of the linguistic expressions;
- for question-answering and language generation systems, the paraphrases allow producing less redundant sentences that show more natural character;
- for textual entailment, the paraphrases allow making inferences on the semantic relations between two statements despite the formal and lexical difference of these statements;
- for semantic interoperability between terminologies, ontologies and textual documents, the paraphrases allow creating links between terms from different semantic resources processed and also between these semantic resources and information contained in documents.

For these different applications, the discovery and acquisition of paraphrases play very important role.

In what follows, we present work on paraphrase (section 1.1) and on paraphrastic reformulation (section 1.2) in linguistics and in NLP (section 1.3). We then specify our objectives (section 1.4).

1.1 Linguistic description of paraphrases

The paraphrase can be described from different points of view. For instance, it can refer to the utterance situation [4–6] and receive contextual values, such as in *two year ago* and *in 2012*. It is then opposed to the linguistic paraphrase, that involves linguistic transformations. Several typologies of linguistic transformations are proposed [7, 8, 2, 9] (the paraphrased elements are underlined):

- morphological paraphrase involves morphological processes (*i.e.*, inflection, affixation and compounding), such as in *We need an improvement of recycling system* and *We need an improved recycling system*;
- lexical paraphrase involves changes at the lexical level with synonyms, hyperonyms, antonyms, etc., such as in *There's a risk of receiving a severe wound* and *There's a possibility of receiving serious injure*;
- semantic paraphrase often covers segments larger than words, such as in *Emma burst into tears* and *Emma cried*;
- syntactic paraphrase reorganizes sentences with the shifting of components or diathesis, such as in *The riddle is solved by him* and *He solved the riddle*, or *Bill sold a car to Tom* and *Bill sold Tom a car*;
- mixed paraphrase may involve various combination of these modifications.

Paraphrase can also be described according to the size of linguistic units involved [10, 11, 2], that distinguishes lexical, sub-phrastic and sentence paraphrases.

The existing classifications of paraphrase are often oriented on one dimension, described with more or less detail (*e.g.* up to 67 lexical functions [7] or 25 categories [9]). As far as we know, the only multidimensional classification considers five dimensions [12]: type of the knowledge required for the production of paraphrases; involved meaning modifications; types of linguistic modifications,

which is close to the classifications above; accuracy of the paraphrastic relation; and mode of production. Besides, paraphrase may also cover two additional dimensions: register of language (*e.g.* specialized *vs.* non-specialized, spoken *vs.* literary [13]); and language (equivalences that correspond to translations [6, 12]). Our acceptance of paraphrase is large but reserved to one language only. Hence, we consider that paraphrase can also be used for description, precision or explanation of ideas previously expressed by a speaker.

1.2 Paraphrastic reformulation

Reformulations occur in formal language and in spoken language, although they show differences [10, 14]. Thus, in oral speech we can observe the elaboration of ideas, that often contains hesitations, false starts, repetitions [15], while in written documents, we find rather its final result [16]. It is usually considered that reformulation is the activity of speakers built on their own linguistic production or on the one of their interlocutor, with or without specific markers. The objective is then to modify some aspects (lexical, syntactic, semantic, pragmatic) but to keep the semantic content constant [17, 18]. Not every reformulation corresponds to paraphrase, and two categories of markers can be thus distinguished (we give examples in French): markers of non-paraphrastic reformulation (*e.g. en somme, en tout cas, de toute façon, enfin*, etc.) and markers of paraphrastic reformulation (called MPRs), like *c'est-à-dire, je m'explique, ça veut dire, en d'autres termes* [19]. With the paraphrastic reformulation, we can distinguish source and target (or paraphrased) entities, usually linked by an MPR. The following criteria are typically used for the distinction of paraphrastic reformulations [19]:

- three phonetic criteria: the repetition of the intonation contour of the sentence; the decrease of the output speed; and a very clear articulation of last syllables at the end of the paraphrase;
- syntactic parallelism of the source and paraphrased entities;
- occurrence of the MPRs, although it is possible to find paraphrases without markers. Among the MPRs, the authors distinguish markers which main task is to establish paraphrastic relations (*e.g. c'est-à-dire*), markers that can establish this relation, and markers that seldom play this role.

The MPRs provide formal mark-up of paraphrastic relations. Notice that the semantic properties of the MPRs allow creating the paraphrase relation among entities that show no semantic equivalence or similarity otherwise [19].

1.3 Natural Language Processing

The two recent literature reviews of methods proposed for the automatic detection of paraphrases [20, 21] state about the increasing importance of this topic for the NLP research. The methods used usually depend on the type of material that is exploited. Often, these methods are based on the paradigmatic properties of linguistic entities and on their capacity to be replaced by each other:

1. *Monolingual corpora.* In monolingual corpora, the string edition similarity [22] and distributional methods are mainly used. In this last case, the linguistic entities (words, phrases, etc) have to share similar vectors to be considered as good candidates for the paraphrase [23, 24];
2. *Monolingual parallel corpora.* When a given text is translated more than once in another language, these translations allow building the monolingual parallel corpus. One of the most used is built with the English translations of *20,000 lieux sous la mer* by Jules Verne. Exploitation of such corpora becomes possible thanks to the methods for word alignment. Various approached are proposed for processing this kind of corpora [25–27];
3. *Monolingual comparable corpora.* Monolingual comparable corpora contain texts on the same event but created independently, like media articles on a given political or social event. The thematic consistency of such texts, the distributional methods and the alignment of comparable sentences allow inducing paraphrastic relations between linguistic entities [28, 29];
4. *Bilingual parallel corpora.* Bilingual parallel corpora, that typically contain translations of a given text in another language, can also be used for the acquisition of paraphrases. In this case, different translations of a given linguistic entity can provide paraphrases [30–33].

1.4 Objectives

Our objective is to work on the detection of paraphrastic reformulations. The originality of our work is related to the following points: (1) The work is done on spoken corpora, that have been very little exploited up to now for the detection of paraphrases [2]; (2) The method for the detection of paraphrastic reformulations is syntagmatic, but not paradigmatic, that is usually dedicated to monolingual corpora [20]; (3) Manual multidimensional annotation of paraphrases is done and provides the reference data; (4) Method for the automatic distinction between paraphrastic and non-paraphrastic reformulations is proposed and tested.

In the following of the paper, we describe the data exploited (section 2) and the method proposed (section 3). We then present and discuss the results (section 4), and outline the directions for future work (section 5).

2 Linguistic data

2.1 Corpora

We use the *ESLO* (Enquêtes SocioLinguistiques à Orléans) corpora [34]: *ESLO1* and *ESLO2*. *ESLO1*, the first sociolinguistic survey in Orléans, France, has been done between 1968 and 1971 by the French department staff from the Essex University, UK in collaboration with the B.E.L.C. (Bureau pour l'étude de l'enseignement de la langue et de la civilisation françaises de Paris) lab. The corpus contains 300 hours of speech, with over 4,500,000 occurrences. The building of the corpus *ESLO2* started in 2008. The objective is to collect over 350

hours of speech with 10 M occurrences. The two corpora are available online[1]. The transcriptions apply two principles: use of the standard spelling and non-use of the written language punctuation. The segmentation is done on *breath groups* detected by the transcribers and on *turns of speech* detectable with the shift of speakers. We use 260 interviews from *ESLO1* (2,349,829 occurrences) and 308 interviews from *ESLO2* (1,412,891 occurrences).

2.2 Markers of paraphrastic reformulation (MPR)

We exploit three MPRs: *c'est-à-dire*, *je veux dire* and *disons*. These MPRs can be translated as *in other words*, *that is to say* and *let's say* respectively. Their common feature is that they are coined on the verb *dire (to say)*. *c'est-à-dire* is the most lexicalized and the most studied [35, 36]: (1) it is used in monologues and dialogues, both spoken and written; (2) the linguistic entities in relation of paraphrase cannot be interchanged because they are not semantically equal; (3) *c'est-à-dire* can shift for instance with *autrement dit* and *en d'autres termes*; (4) it creates paraphrase relation between entities without semantic equivalence; (5) in addition to the three prototypical functions (correction, reformulation and argumentation) it can also mark conclusion, justification and hesitation. Concerning *disons*, its known characteristics are [37]: (1) it is semantically close to *je veux dire*; (2) *disons* and *eh bien* present analogy because they mark the break between two utterances; (3) *disons* and *enfin* present analogy in correction contexts. It is impossible to remove *disons* because the target entity conveys different semantics [38]. Finally, *je veux dire* is known to have several meanings and can be replaced by *autrement dit, c'est-à-dire* [39]. The markers studied can have several functions. We use them for their paraphrastic reformulation function.

3 Methodology for the detection of paraphrases

Utterances that contain one of the MPRs studied are extracted from the corpora and pre-processed (section 3.1). The method relies on manual (section 3.2) and automatic (section 3.3) processing of corpora. The analysis and evaluation of the results is done (section 3.4).

3.1 Pre-processing of corpora

In order to rebuild the utterances, the transcription files are segmented in turns of speech: new utterance begins with the shift of speakers. In case of speaker overlapping, the overlapped segments are associated with all the involved speakers. The corpora are then POS-tagged and analysed with the SEM chunker [40] adapted to spoken language. SEM detects minimal chunks.

[1] *http://eslo.tge-adonis.fr/*

3.2 Manual annotation of paraphrastic reformulations

The manual annotation allows first distinguishing between paraphrastic and non-paraphrastic reformulations. For paraphrastic reformulations, the annotation is finer-grained. It applies to the source and target entities related by the MPRs, and to the paraphrastic relation. The annotation is done along several dimensions, some of which are inspired by the existing classifications (section 1):

1. *Syntactic tag*: each entity is annotated with its POS-tag (*e.g.* N, A, V, Prep) or syntactic constituent (*e.g.* NP, VP, AP, PP). Size of entities is defined according to the semantics of the paraphrase, but not on the basis of chunks;
2. Each relation is annotated with:
   - *rel-lex*: type of lexical relation among the two paraphrased entities (*e.g.* hyperonym, synonym, antonym, instance, meronym);
   - *modif-lex*: type of lexical modification (*e.g.* remplacement, deletion, insertion);
   - *modif-morph*: type of morphological modification (*i.e.* inflection, derivation or compounding);
   - *modif-synt*: type of syntactic modification (*e.g.* active/passive);
   - *rel-pragm*: type of pragmatic relation, linked to the function of paraphrase and reformulation, inspired by the existing typologies [17, 18]. We distinguish: definition, explanation, exemplification, precision, denomination, result, linguistic or referential correction, and equivalence.

Annotation examples can be found in (1) and (2): annotation is in gray, the file reference between brackets. We can see for instance that entities {*Saint Jean de la Ruelle*, *Orléans*} in (1) and {*démocratiser l'enseignement (democratize the education)*, *permettre à tout le monde de rentrer en faculté (allow everybody to enter the university)*} in (2) have the paraphrase relation.

(1) *pendant nous avons fait grève à la Régie Renault euh de* <NP1>*Saint Jean de la Ruelle*</NP1> <MPR>*c'est-à-dire*</MPR> <NP2 rel-lex="mer (Saint Jean de la Ruelle/Orléans)" rel-pragm="cor-ref">*Orléans*</NP2> *parce que c'est ça fait partie d'Orléans* [ESLO1_ENT_149_C]

(2) <VP1>*démocratiser l'enseignement*</VP1> <MPR>*c'est-à-dire* </MPR> <VP2 rel-lex="syn(démocratiser/permettre à tout le monde) syn(enseignement/faculté)" modif-lex="ajout(rentrer à)" rel-pragm="explic">*permettre à tout le monde de rentrer en faculté*</VP2> [ESLO1_ENT_121_C]

3.3 Automatic detection of paraphrastic reformulations

The main objective is to decide whether a given occurrence of MPR creates the paraphrastic reformulation relation or not. Several filters are applied for this:

- if the MPR is at the beginning or end of utterance, the context is not sufficient to create paraphrastic relation;
- if the MPR is found in specific lexical contexts, such as occurrence of *nous* with *disons (we say)*, we consider that such contexts are not paraphrastic;

- if the MPR occurs with other repeated discursive markers (*donc, enfin, quoi*), hesitation euh, interjections (*ben hm ouais*), primes (*s-*), etc., we consider that the MPR is part of oral disfluencies [15] and is not paraphrastic;
- if the MPR occurs within expression or phrase, like *indépendamment de (independently of)* in example (3), we consider that the context is not paraphrastic. This test is done with the syntactically chunked output. In order to verify whether the expression or phrase exist, we query an online search engine and analyse the frequencies attensted on the web. We assume that the web frequencies provide with information that is more exhaustive than frequencies found in reference corpora. Each segment is tested in three ways: with one, two or three chunks on the right and on the left of the MPR, excepting the disfluence markers. Size of the tested segments is empirically set to seven words at most. Then, we compute the average frequency for the three kinds of segments (one, two or three chunks on the right and on the left of the MPR). The average frequency of the segments must not be lower than the threshold tested, that is between 10 and 6,000. If the average frequency is higher than the threshold, the test indicates that the expression or phrase exist in the language and that the MPR represents the disfluency.

3.4 Analysis and evaluation

The annotation protocol has been fixed on a subset of *ESLO1*, while the evaluation is done on the remaining *ESLO1* subset and on the *interviews* from *ESLO2*. Two kinds of evaluation are performed: (1) manual annotation is checked for the inter-annotator agreement at the level of the paraphrastic relation. With two sets of annotations, we apply the Cohen kappa [41] measure; (2) precision of automatic detection of the paraphrastic relation is evaluated against the manual annotation. The analysis of results addresses the frequency of relations and their attributes. We are particularly interested in the existence of paraphrastic relations, in the syntactic equivalence between the entities and the existence of morphological modifications (inflection, derivation or compounding), that can give formal indications on the paraphrase.

4 Results and Discussion

4.1 Building and pre-processing of corpora

In Table 1, we indicate the size of corpora in number of words, the number and average size of turns of speech, the number of utterances with the MPRs studied, and the size of these utterances. the average size of turns of speech is between 14 and 19 words, with many minimal utterances (one or two words). *c'est-à-dire* is the most frequent, as it provides over half of utterances. *disons* is particularly frequent in *ESLO1*, but the less frequent in *ESLO2*. The difference may be due to the diachronic evolution, as other words may have taken the corresponding discursive function. Concerning the average size of utterances with MPRs, it

Table 1. Description of corpora: size, number and average size of turns of speech, number of utterances with three MPRs studied, size of utterances with the MPRs.

	ESLO1	*ESLO2*
number of transcription files	260	308
size of corpora (occ of words)	2,349,829	1,412,891
average size of transcription files	9,037,80	4,587,31
number of turns of speech	166,602	70,707
average size of turns of speech	14.10	19.98
c'est-à-dire	1,849	594
je veux dire	285	291
disons	1,068	183
total number of utterances with MPRs	3,202	1,068
size of utterances with MPRs (minimal)	1	1
size of utterances with MPRs (maximal)	6,382	1,050
size of utterances with MPRs (average)	62.88	86.34

Table 2. Jugment on the paraphrastic relation for the two annotators A1 and A2.

	ESLO1					*ESLO2*				
	A1		*A2*		*agr.*	*A1*		*A2*		*agr.*
	yes	*no*	*yes*	*no*		*yes*	*no*	*yes*	*no*	
c'est-à-dire (number)	96	193	66	223	249	74	124	65	137	162
je veux dire (number)	16	49	8	57	57	47	91	27	110	107
disons (number)	18	104	8	115	106	10	45	9	46	46
total utterances with MPRs (number)	130	346	82	395	412	131	260	101	293	315
total utterances with MPRs (%)	27	73	17	83		33	67	26	74	

is quite high (62.88 in *ESLO1* and 86.34 in *ESLO2*). We assume that these utterances can contain paraphrases and show the genesis of speaker ideas [16, 15]. This can also explain the fact that the average size of these utterances is higher than the global average observed in corpora. We can observe that the maximal size of utterances is very high and can reach up to 1,050 in *ESLO2* and 6,382 in *ESLO1*.

4.2 Manual annotation of paraphrases

We annotated 476 utterances in *ESLO1* and 394 utterances in *ESLO2* (54 and 30 interviews respectively) that contain the MPRs. These annotations are our reference data. Table 2 indicates the annotation results provided by the two annotators. The annotators state that between 17 and 27% of utterances are paraphrastic in *ESLO1*, and between 26 et 33% in *ESLO2*. Annotator *A1* accepts more contexts as paraphrastic. The inter-annotator agreement is substantial (0.617) in *ESLO1* and moderate (0.526) in *ESLO2*. This is a good agreement given the inherent subjectivity induced by these data. As observed in the literature, these MPRs can occur in paraphrastic and non-paraphrastic contexts. In

Table 3. Percentage of paraphrastic and non-paraphrastic constructions with MPRs.

	ESLO1				*ESLO2*			
	A1		*A2*		*A1*		*A2*	
	yes	*no*	*yes*	*no*	*yes*	*no*	*yes*	*no*
c'est-à-dire (%)	33	67	22	78	37	63	32	68
je veux dire (%)	25	75	12	88	34	66	20	80
disons (%)	15	85	7	93	18	82	6	94

example (3), that do not contain paraphrases, the MPR is to be associated with discursive markers and disfluencies.

(3) *différence sensible entre vos différents clients dans leur façon de choisir la viande dans ce qu'ils achètent et caetera indépendamment* <MPR>*disons* </MPR> *de leurs oui origines de classe* [ESLO1_ENT_001_C]

Table 3 indicates the percentage of paraphrastic and non-paraphrastic constructions with MPRs. The two annotators consider that *c'est-à-dire* is the most grammaticalized in this function because it introduces the largest number of paraphrases, while *disons* is the less grammaticalized. Concerning *disons*, we assume that it is ambiguous: in addition to the paraphrase, it can also mean *dire (to say)* and show discursive [17] or disfluency function (example (3)).

In over 70% of contexts, there is no syntactic equivalence between the entities in paraphrastic relations (examples (4) and (5)). This aspect depends on the annotator choice: for instance, in (4), various segments can be selected (*les gens me semblent plus plus affables*, *plus affables* or *affables*). Another interesting fact is related to the morphological modifications: only ten such modifications are observed in each corpus (*e.g.* {*achat (purchase)*, *achète (buy)*}, {*connais (know)*, *connu (known)*}, {*pourrait (could)*, *pouviez (can)*}, {*client (client)*, *clientèle (clientele)*}, {*manoeuvres (manoeuvres)*, *manuel (manual)*}). This means that very few formal cues are available for the detection of paraphrases. Besides, we find only one occurrence with syntactic modifications (active/passive). Concerning lexical modifications, we mainly observe replacements. As noticed in the literature [19], we can find several paraphrastic reformulations in which entities have no semantic relation except the one marked by the MPR, such as {*conférences (conferences)*, *causeries (chat)*} in (6).

(4) *je préfère mieux le le nord de la France franchement le département du Nord et le département du Pas-de-Calais où* <P1>*les gens me semblent plus plus affables*</P1> <MPR>*disons*</MPR> *euh* <PP2 rel-lex="syn" rel-pragm="explic">*avec qui j'ai on a plus facilement des des rapports agréables*</PP2> [ESLO1_ENT_003_C]

(5) *y a le euh le le plus grand goup- groupe et puis euh ce qu'on appelle* <NP1>*toujours les mêmes*</NP1> <MPR>*c'est-à-dire*</MPR> *euh* <P2

Table 4. Precision of the automatic detection of paraphrastic reformulations.

	ESLO1		*ESLO2*	
	A1	A2	A1	A2
lexical and discursive filters	40.5	40.5	37.7	37.8
lexical and discursive filters + frequency (>6000)	25.8	25.9	18.7	18.9
lexical and discursive filters + priority frequency (>6000)	63.0	63.0	66.4	66.3

rel-lex="syn" rel-pragm="equiv">tous ceux qu'on connait</P2> quoi [ESLO2_ENT_1004_C]

(6) *des conférences y en a assez souvent sur France culture enfin <MPR>disons</MPR> des causeries* [ESLO1_ENT_121_C]

Among the lexical relations, synonymy and hyperonymy are the most frequent, followed by instances with named entities, equivalence and result. According to the pragmatic relations, we can distinguish three functions of MPRs:

- possibility to add new information with explanation, precision, exemplification and definition. This function can be associated with the known functions (correction, reformulation and argumentation [35]). In these situations, the target entity is richer and clairer, like in examples (4) and (2);
- possibility to tell the same thing, but using other linguistic means with the equivalence relation, such as in (5) and (6). We consider that, contrary to what has been noticed in the literature [38], with these relations the source and target entities are exchangeable and we can remove the MPR;
- with the relation *result*, we can observe inverse situation to the explanation: the target entity can be shorter than the source entity (example (7)).

(7) *voilà <P1>le côté très betonné voilà c'est pas ils ont pas dveloppé les les logements étudiants suffisamment ils ont pas développé l'off- l'offre culturelle euh en même temps</P1> donc enfin <MPR>je veux dire</MPR> voilà <P2 rel-pragm="res">c'est mort</P2>* [ESLO2_ENT_1012_C]

4.3 Automatic detection of paraphrastic reformulations

Precision of the automatic detection of paraphrastic reformulations is indicated in Table 4. The results are coherent between the two annotators in the two corpora processed, although it is more complicated to correctly process the *ESLO2* corpus. The lexical and discursive filters reach 40% and 38% precision in *ESLO1* and *ESLO2* respectively. The additional use of the frequency filters decreases the results to 26% and 19%. But, when the frequency filters have priority on the lexical and discursive filters, we improve precision to up to 63% and 66%: in this case, we consider that frequency is indicative of the paraphrases even if the utterance contains oral disfluencies. Notice that precision is improved with with the increasing of the threshold. The highest threshold tested is 6,000, while the

improvement of precision is observed with the average frequency between 10 and 4,500. Above that threshold, we observe no evolution of the precision values.

The precision we obtain can be considered as acceptable. It is comparable or even superior to the precision obtained in previous work [2]. By comparison with paraphrase recognition results obtained on the MSR written corpus, that is annotated mainly with lexical, syntactic and contextual paraphrases [9], our results are similar to those provided by the baselines and some of the systems reported [21]. We expect to reach better results in the next future.

5 Conclusion and Future work

We have proposed a method for the detection of paraphrastic reformulations in monolingual spoken corpora in French (*ESLO1* and *ESLO2*). One originality is that we take into account the specificity of the spoken data through the building of utterances, the consideration of oral disfluencies, and the use of the NLP tool adapted to spoken corpora [40]. Another originality is that we address the detection of paraphrastic reformulations with syntagmatic approach, while usually paradigmatic approaches are used with this kind of data. We accept a large acceptance of paraphrase [7, 9], that also covers clarification, explanation, or synthesis of ideas uttered previously by a speaker. We perform manual annotation and automatic detection of paraphrases. The manual multidimensional annotation allows producing the reference data and observations on the paraphrase relations in spoken corpora. These data allow evaluating the automatic method. The inter-annotator agreement is 0.617 and 0.526 in *ESLO1* and *ESLO2* respectively. The automatic recognition of paraphrases relies on a set of filters (lexical, discursive and frequency) and reaches up to 66.4%. The comparison with the existing work confirms some previous observations [19]: (1) reformulations are not always paraphrastic, and can perform other functions; (2) MPRs can create paraphrastic relations between entities that do not show semantic equivalence otherwise. On contrary, we seldom observe syntactic equivalence between source and target entities, we assume that it is possible to exchange places of entities with the equivalence relation, and that it is possible to remove the MPR.

We have several directions for future work. We plan to involve additional annotators and organize conciliation meetings to obtain more consensual reference data. Other MPRs can be studied and compared among them. For the automatic detection of paraphrastic reformulations, we can improve the current performance thanks to a better recognition of repetitions and to machine learning. The automatic detection of boundaries of source and target entities in another perspective. We plan also to compare the paraphrastic reformulations in spoken and written corpora: we assume the process is similar, as it allows making ideas clearer, and dissimilar from the cognitive point of view [16, 15]. As indicated, the two corpora exploited have been built with similar principles but with 40 year difference. This offers the possibility to perform diachronic study of MPRs. Besides, a similar study can be done on corpora from other languages. We can also combine this study with the social data on speakers.

References

1. François, F.: La communication inégale. Heurs et malheurs de linteraction verbale. In: Actualités pédagogiques et psychologiques. Delachaux & Niestlé, Neuchâtel-Paris (1990)
2. Bouamor, H., Max, A., Vilnat, A.: Étude bilingue de l'acquisition et de la validation automatiques de paraphrases sous-phrastiques. TAL **53**(1) (2012) 11–37
3. Boucheron, S.: La langue de l'un, et celle de l'autre: l'entre parenthèses comme aire de reformulation. In: Répétition, Altération, Reformulation. Presses Universitaires Franc-Comtoises, Besançon (2000) 113–118
4. Martin, R.: Inférence, antonymie et paraphrase. Klincksieck, Paris (1976)
5. Vezin, L.: Les paraphrases: étude sémantique, leur rôle dans l'apprentissage. L'année psychologique **76**(1) (1976) 177–197
6. Fuchs, C.: Paraphrase et énonciation. Orphys, Paris (1994)
7. Melčuk, I.: Paraphrase et lexique dans la théorie linguistique sens-texte in lexique et paraphrase. Lexique **6** (1988) 13–54
8. Vila, M., Antònia Mart, M., Rodríguez, H.: Paraphrase concept and typology. a linguistically based and computationally oriented approach. Procesamiento del Lenguaje Natural **46** (2011) 83–90
9. Bhagat, R., Hovy, E.: What is a paraphrase? Computational Linguistics **39**(3) (2013) 463–472
10. Flottum, K.: Dire et redire. La reformulation introduite par "c'est-à-dire". Thèse de doctorat, Hogskolen i Stavanger, Stavanger (1995)
11. Fujita, A.: Typology of paraphrases and approaches to compute them. In: CBA to Paraphrasing & Nominalization, Barcelona, Spain (2010) Invited talk.
12. Milicevic, J.: La paraphrase : Modélisation de la paraphrase langagière. Peter Lang (2007)
13. Elhadad, N., Sutaria, K.: Mining a lexicon of technical terms and lay equivalents. In: BioNLP. (2007) 49–56
14. Rossari, C.: De l'exploitation de quelques connecteurs reformulatifs dans la gestion des articulations discursives. Pratiques **75** (1992) 111–124
15. Blanche-Benveniste, C., Bilger, M., Rouget, C., Van Den Eynde, K.: Le français parlé. Études grammaticales. CNRS Éditions, Paris (1991)
16. Hagège, C.: L'homme de paroles. Contribution linguistique aux sciences humaines. Fayard, Paris (1985)
17. Gülich, E., Kotschi, T.: Les actes de reformulation dans la consultation La dame de Caluire. In Bange, P., ed.: L'analyse des interactions verbales. La dame de Caluire: une consultation. P Lang, Berne (1987) 15–81
18. Kanaan, L.: Reformulations, contacts de langues et compétence de communication: analyse linguistique et interactionnelle dans des discussions entre jeunes Libanais francophones. Thèse de doctorat, Université d'Orléans, Orléans (2011)
19. Rossari, C. In: Les oprations de reformulation. Analyse du processus et des marques dans une perspective contrastive franais-italien. (1993)
20. Madnani, N., Dorr, B.J.: Generating phrasal and sentential paraphrases: A survey of data-driven methods. Computational Linguistics **36** (2010) 341–387
21. Androutsopoulos, I., Malakasiotis, P.: A survey of paraphrasing and textual entailment methods. Journal of Artificial Intelligence Research **38** (2010) 135–187
22. Malakasiotis, P., Androutsopoulos, I.: Learning textual entailment using SVMs and string similarity measures. In: ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. (2007) 42–47

23. Lin, D., Pantel, L.: Dirt - discovery of inference rules from text. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining. (2001) 323–328
24. Pasa, M., Dienes, P.: Aligning needles in a haystack: Paraphrase acquisition across the Web. In: IJCNLP. (2005) 119–130
25. Barzilay, R., McKeown, L.: Extracting paraphrases from a parallel corpus. In: ACL. (2001) 50–57
26. Ibrahim, A., Katz, B., Lin, J.: Extracting structural paraphrases from aligned monolingual corpora. In: International Workshop on Paraphrasing. (2003) 57–64
27. Quirk, C., Brockett, C., Dolan, W.: Monolingual machine translation for paraphrase generation. In: EMNLP. (2004) 142–149
28. Shinyama, Y., Sekine, S., Sudo, K., Grishman, R.: Automatic paraphrase acquisition from news articles. In: Proceedings of HLT. (2002) 313–318
29. Sekine, S.: Automatic paraphrase discovery based on context and keywords between NE pairs. In: International Workshop on Paraphrasing. (2005) 80–87
30. Bannard, C., Callison-Burch, C.: Paraphrasing with bilingual parallel corpora. In: ACL. (2005) 597–604
31. Madnani, N., Resnik, P., Dorr, B., Schwartz, R.: Applying automatically generated semantic knowledge: A case study in machine translation. In: NSF Symposium on Semantic Knowledge Discovery, Organization and Use. (2008) 60–61
32. Callison-Burch, C., Cohn, T., Lapata, M.: Parametric: An automatic evaluation metric for paraphrasing. In: COLING. (2008) 97–104
33. Kok, S., Brockett, C.: Hitting the right paraphrases in good time. In: NAACL. (2010) 145–153
34. Eshkol-Taravella, I., Baude, O., Maurel, D., Hriba, L., Dugua, C., Tellier, I.: Un grand corpus oral "disponible" : le corpus d'Orléans 1968-2012. Traitement Automatique de Langues **52**(3) (2012) 17–46
35. Hölker, K.: Zur Analyse von Markern. Franz Steiner, Stuttgart (1988)
36. Beeching, K.: La co-variation des marqueurs discursifs bon, c'est-à-dire, enfin, hein, quand même, quoi et si vous voulez : une question d'identité ? Langue française **154**(2) (2007) 78–93
37. Hwang, Y.: Eh bien, alors, enfin et disons en français parlé contemporain. L'Information Grammaticale **57** (1993) 46–48
38. Petit, M.: Discrimination prosodique et représentation du lexique : application aux emplois des connecteurs discursifs. Thèse de doctorat, Université d'Orléans, Orléans (2009)
39. Teston-Bonnard, S.: Je veux dire est-il toujours une marque de reformulation? In Bot, M.L., Schuwer, M., Richard, E., eds.: Rivages linguistiques. La Reformulation. Marqueurs linguistiques. Stratégies énonciatives. PUR, Rennes (2008) 51–69
40. Dupont, Y., Tellier, I., Courmet, A.: Un segmenteur-étiqueteur et un chunker pour le français. Technical report, LIFO, Université d'Orléans (2012) demo.
41. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**(1) (1960) 37–46