Disambiguation of occurrences of reformulation markers c'est-à-dire, disons, ça veut dire

Natalia Grabar, Iris Eshkol-Taravella

Disambiguation of occurrences of reformulation markers *c'est-à-dire, disons, ça veut dire*

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

[1]UMR8163 STL CNRS, Université Lille 3 – France

[2]UMR7270 LLL CNRS, Université d'Orléans – France

Abstract

Reformulation is a process which consists of saying again an utterance which has already been said, but which goes through formal and/or semantic modifications. Sometimes, reformulations are signaled by specific markers, such as *c'est-à-dire, disons, ça veut dire.* We propose to study the reformulation phenomenon. More particularly, we concentrate on the syntagmatic structure *S1 marker S2*, coined around the reformulation markers, and in which the first segment *S1* is reformulated by the second segment *S2*. The purpose of our study is to automatically differentiate between reformulation and non-reformulation occurrences of the markers studied. We design a rule-based system which relies on a set of rules to make the decision. Two kinds of French corpora are processed: spoken corpora ESLO and forum discussion corpus. The evaluation of the system is performed against the manually annotated and consensual reference data. Our system has been created on a subset of the spoken corpus and then applied to the rest of the data. The results obtained reach up to 0.75 precision and are comparable on the corpora analyzed, although spoken corpora remain more difficult to process.

Résumé

La reformulation est un processus qui consiste à dire à nouveau une information qui a déjà été dite, mais en effectuant un ensemble de modifications formelles et/ou sémantiques. Parfois, les reformulations sont signalées par des marquers spécifiques, comme par exemple *c'est-à-dire, disons, ça veut dire.* Nous proposons d'étudier le phénomène de reformulation. Plus particulièrement, nous nous concentrons sur la structure syntagmatique *S1 marker S2*, formée autour d'un marqueur de reformulation, et dans laquelle le premier segment *S1* est reformulé par le deuxième segment *S2*. L'objectif de notre étude est de différencier automatiquement les occurrences reformulatives et non reformulatives des marqueurs étudiés. Nous créons un système basé sur des règles, qui repose sur un ensemble d'indices pour prendre la décision. Deux types de corpus en français sont traités : corpus oral ESLO et corpus de discussion de forum. L'évaluation du système est effectuée grâce à une comparaison avec un ensemble de référence consensuel annoté manuellement. Notre système a été créé sur un sous-ensemble du corpus oral et ensuite appliqué au reste de corpus. Les résultats obtenus atteignent jusqu'à 0,75 de précision et sont comparables dans les corpus analysés, bien que les corpus oraux soient plus difficiles à traiter.

Key words: reformulation, spoken and written corpora, automatic detection of reformulation

1. Introduction

Reformulation is a process which consists of saying again and in different way an utterance which has already been said (Le Bot et al., 2008). Reformulation can be performed by demand of the interlocutor, or by the decision of the speaker himself. Several reasons may lead to the reformulation: make previous statement more precise, explain the previous statement, give a name to what has been described, etc. Reformulation is spread in written and spoken, formal and informal languages, although, in all these situations, reformulation shows

some specificities (Fløttum, 1995; Rossari, 1992). Thus, in written texts, reformulation corresponds to the final result which is proposed to the user, while in spoken language, reformulation represents the various steps of elaboration of an idea (Levelt, 1983; Hagège, 1985; Blanche-Benveniste et al., 1991; Blanche-Benveniste, 1995). Reformulation can be marked by specific markers *(c'est-à-dire, autrement dit, disons, je m'explique, ça veut dire, en d'autres termes...*), such as in

(1) ***je suis bien*** *je veux dire* ***je me sens bien***

but it can also occur without such markers, in which case, phonetic, prosodic and semantic layers of the language may be indicative of its occurrence:

(2) *même en bouquin enfin* ***je suis très polar j'aime beaucoup les polars***

The proposed work is done on corpora in French. It is dedicated to reformulations introduced by three markers: *c'est-à-dire, disons, ça veut dire*. The research problem we propose to address is related to the fact that these markers do not systematically indicate reformulations, but can also show other kinds of occurrences. Let us observe the examples which follow and in which the occurrences of *disons* are used with different meanings:

(3) *il y a énormément de vieilles familles euh bourgeoises ...* ***ils sont souvent disons moins aisés*** *que les familles d'ouvriers les familles d'employés*

(4) *nous avons une expression chez nous ...* ***nous disons*** *que les gens qui gardent ces choses ont euh une mentalité d'écureuil*

(5) ***basée*** *euh* ***sur le capitalisme*** *enfin la société française* ***disons*** *euh euh* ***basée sur les valeurs*** *euh euh* ***erronées***

In the first example, *disons* is used as discursive marker. It occurs within the syntactic group *ils sont moins aisés (they are not very rich)*. In the second example, the occurrence of *disons* is in fact the inflectional form (plural of the first person, present tense) of the verb *dire* (*to say*). Only the third example provides the occurrence of *disons* used as reformulation marker.

The objective of the proposed work is to detect sentences with reformulations coined around the three markers studied (*c'est-à-dire, disons, ça veut dire*). The task aims at deciding whether a given occurrence of a marker introduces reformulation or not. This task started to be studied in a previous work (Eshkol-Taravella & Grabar, 2014). By comparison with this work, the reference data are consensual: the annotations have been done by two annotators and then, during common work sessions, a consensus has been reached on annotations which differed from one annotator to another. This process permits to build one common set of annotated data instead of the two sets annotated. Besides, we extended the types of the corpora studied: we work now with two kinds of corpora (spoken and written corpora) instead of only spoken corpora exploited in previous work.

In what follows, we will first present related work on close topics (Section 2). We then introduce the linguistic data processed (Section 3), and the method defined for performing the task (Section 4). Then, the results are presented and discussed (Section 5). We conclude with directions for future work (Section 6).

2. Related Work

Reformulation can be described from different points of view. We should notice first that reformulation can be conceived as paraphrastic variation of a linguistic segment in which

formal modifications occurred (Neveu, 2004). In this case, paraphrase is the result of reformulation. These two notions are thus closely related. Still, in the existing work, these two notions remain separated. Thus, several works have been done on paraphrases: their automatic detection (Malakasiotis & Androutsopoulos, 2007; Lin & Pantel, 2001; Ibrahim et al., 2003; Sekine, 2005; Kok & Brockett, 2010) and description, which we present below.

For the description and classification of paraphrases and reformulations, several points of view are possible. Thus, they can refer to the utterance situation (Martin, 1976; Culioli, 1987; Vezin, 1976; Fuchs, 1994; Vion , 2006) and receive contextual values, such as in *two year ago* and *in 2014*. They are then opposed to linguistic transformations within reformulated and paraphrased units. Several typologies of linguistic transformations are proposed (Vila et al., 2011; Bouamor et al., 2012; Bhagat and Hovy, 2013) (the reformulated 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;*
- mixed paraphrase may involve various combination of these modifications.

In the Sens-Texte theory (Melčuk, 1988), a set of lexical functions is proposed. Their purpose is to describe and codify lexical and semantic relations between linguistic elements. The general objective of these functions is to describe transformation rules of a given language. Some of these functions (*eg, Syn, Conv, Contr, Anti, Gener, Sing, Mult, Cap, Equip*) can be exploited for the encoding of paraphrases.

Paraphrase and reformulation can also be described according to the size of linguistic units involved (Fløttum, 1995; Fujita, 2010; Bouamor et al., 2012), that distinguishes lexical, sub-phrastic and sentence segments.

Notice that there are also several existing classifications of paraphrase, described with more or less detail: *e.g.* up to 67 lexical functions (Melčuk, 1988) or 25 categories (Bhagat et al., 2013).

As for the reformulation, this research point has been mainly addressed through the study of reformulation markers: *c'est-à-dire* (Gulich & Kotschi, 1983; Roulet, 1987; Holker, 1988; Beeching, 2007), *je veux dire* (Teston-Bonnard, 2008) and *disons* (Hwang, 1993; Petit, 2009; Saunier, 2012). The common issue for these markers is that they are all coined on the same verb *dire (to say)*. It is also recognized that *c'est-à-dire* is the most lexicalized and the most studied within this set of markers. All of them have been recognized to introduce the reformulations, called paraphrastic reformulations in the existing work. But they can also play other roles in the discourse, such as argumentation or spoken disfluencies.

In our work, we make a large acceptation of the reformulation (and paraphrase) phenomena. For its description, we propose a specific multi-dimensional annotation schema presented in Section 4.2.

3. Linguistic Data Processed

We use two kinds of corpora: two spoken corpora *ESLO* (Section 3.1) and a corpus with forum discussions (Section 3.2).

3.1. ESLO corpus

We use the *ESLO* (*Enquêtes SocioLinguistiques à Orléans*) corpora (Eshkol et al., 2012): *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. The corpora provide different genres, such as meetings, interviews, shop and school discussions. In order to study comparable data, we use 260 interviews from *ESLO1* (2,349,829 occurrences) and 308 interviews from *ESLO2* (1,412,891 occurrences).

3.2. Forum corpus

The forum corpus is collected from the discussion forum *Hypertension* from Doctissimo[2]. This corpus provides 12,588 threads with 67,652 messages, and 6,788,361 word occurrences. Messages are written by the internet users, who need to speak about their illness and life. By comparison with standard written texts, forum discussions are non-normed writings, which can contain misspellings, syntax errors, and other non-conventional linguistic items (specific abbreviations, emoticons...). Besides, the forum discussions also present the specificities of the spoken language (absence of standard punctuation, frequent reformulations and disfluencies, primes, etc.). Forum discussions may be conceived as a hybrid form of spoken and written language.

4. Methods Proposed for Filtering the Data

Utterances that contain one of the markers studied are extracted from the corpora and pre-processed (Section 4.1). Then, the method relies on manual (Section 4.2) and automatic

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

[2] http://forum.doctissimo.fr/sante/hypertension-problemes-cardiaques/liste_sujet-1.htm (collected in May 2013)

(Section 4.3) processing of the data. The analysis and evaluation of the results is performed in the last sub-section (Section 4.4).

4.1. Pre-processing of Corpora

The pre-processing step varies according to the corpora processed.

In the spoken corpora *ESLO,* 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. If the speaker continues the speech after the overlapping, his turn of speech continues as well. The spoken corpora are then POS-tagged and analyzed with the SEM chunker (Dupont et al., 2012) adapted to spoken language. This chunker is probabilistic and has been trained on specifically annotated spoken corpus. The, the generated model can be applied to other spoken corpora in French. The chunker first performs the syntactic tagging of the input text which allows to associate word occurrences with their syntactic categories. For instance in the sequence *nous avons fait grève, nous* is a personal pronoun, *avons* is a conjugated form of the auxiliary verb *avoir*, *fait* is a participial form of the verb *faire*, and *grève* is a noun. Then, the chunker performs the shallow syntactic analysis for grouping the words within common chunks, which is the smallest sequence of linguistic entities forming a syntactic group with one syntactic head. SEM detects minimal chunks, such as (*nous avons) (fait grève)*.

The forum corpus is POS-tagged and syntactically analyzed with Cordial (Laurent et al., 2009). Cordial provides information similar to SEM: POS-tags and syntactic groups, but also the lemmas of word forms. Cordial is rule-based and exploits a set of syntactic rules for the definition of POS-tags and lemmas within sentences, and for their syntactic parsing.

Our study concentrates on sentences and enunciations which contain the markers studied.

4.2. Manual Annotation of Reformulations

The manual annotation is performed at two levels: detection of the reformulations, and a fine-grained annotation of these reformulations. The annotation applies to the source *S1* and target *S2* segments (or entities) related by the markers, and to the reformulation relation. The annotation is done along several dimensions, some of which are inspired by the existing classifications:

- *Syntactic tag*: each entity is annotated with its POS-tag (e.g. N for noun, A for adjective, V for verb, Prep for preposition) or syntactic constituent (e.g., NP for noun phrase, VP for verbal phrase, AP for adjectival phrase, PP for prepositional phrase). Size of entities is defined according to the semantics of the reformulation, but not on the basis of chunks (in spoken corpora) or syntactic groups (in written corpus).

Each reformulation relation is then annotated with:

- *rel-lex*: type of lexical relation among the two segments (e.g., hyperonym, synonym, antonym, instance, meronym);
- *modif-lex*: type of lexical modification (i.e. replacement, 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 (Gulich & Kotschi, 1987; Kanaan, 2011). We distinguish eleven categories: definition, explanation, exemplification, precision, justification, denomination, result, linguistic or referential correction, opposition, and paraphrase (or equivalence).

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

(6) *pendant nous avons fait grève à la Régie Renault euh de <NP1>Saint Jean de la Ruelle</NP1> <MR>c'est-à-dire</MR> <NP2 rel-lex="mero(Saint Jean de la Ruelle/Orléans)" rel-pragm="cor-ref">Orléans</NP2> parce que c' est ça fait partie d' Orléans* [ESLO1_ENT_149]

(7) *euh <VP1>démocratiser l'enseignement</VP1> <MR>c'est-à-dire</MR> <VP2 rel-lex="syno (démocratiser/permettre à tout le monde) syno(enseignement/faculté)" modif-lex="ajout(rentrer à)" rel-pragm="explic">permettre à tout le monde de rentrer en faculté</VP2>* [ESLO1_ENT_121]

In Table 1, we indicate the size of the annotated reference data which are exploited in this study for observations, for fitting of the automatic system and for the evaluation of the system. These data only contain sentences and enunciations with the three reformulation markers studied (*c'est-à-dire, disons, ça veut dire*).

Corpus	*Number of enunciations/sentences*	*Number of occurrences of words*
ESLO1	477	19,832
ESLO2	394	28,945
Forum	193	9,194
Total	1,064	57,971

Table 1. Size of the annotated data: sentences and enunciations with the three reformulation markers studied.

4.3. Automatic Detection of Reformulations

The purpose of the automatic processing proposed is to decide whether a given occurrence of marker is related to reformulation or not. On the whole set of occurrences of markers in the corpora processed, we apply several filters:

- if the marker is at the beginning or end of an utterance or sentence, we consider that the context is not sufficient to establish reformulation relation;
- if the marker is found in specific lexical contexts, such as occurrence of *nous* with *disons* (*we say)*, we consider that such contexts are not related to reformulations;
- if the marker occurs with other repeated discursive markers (*donc, enfin, quoi…*), hesitation euh, interjections (*en hm ouais*), primes (*s-*), etc., we consider that the marker is part of spoken disfluencies (Blanche-Benveniste et al., 1991) and is not related to reformulations;
- if the marker occurs within existing expression or phrase, like *indépendamment de* (*independently on*) in the example below (8), we consider that the context does not contain reformulation. This test is done on the syntactically analyzed and chunked output. In order to verify whether the expression or phrase exist in language, we query an online search engine and analyze the frequencies attested on the web. We assume that the web frequencies provide with information that is more exhaustive than frequencies found in the French reference corpora Frantex (Quemada, 1992). Thus, each segment is tested in three ways: with one, two or three chunks or syntactic groups on the right and on the left of the marker, 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 or syntactic groups on the right and on the left of the marker). The average frequency of the segments must not be lower than the thresholds tested, that are between 10 and 6,000. If the average frequency is higher than the threshold, the test indicates that the expression or phrase exist in language and that the marker represents the disfluency. Thus, the occurrence is not related to reformulations.

  (8) *est-ce que vous remarquez une différence sensible entre vos différents clients dans leur façon de choisir la viande dans ce qu'ils achètent et caetera* ***indépendamment*** *<MR>disons</MR>* ***de*** *leurs oui origines de classe [ESLO1_ENT_001]*

4.4. Analysis and Evaluation

The annotation protocol has been fixed on a subset of *ESLO1,* while the evaluation is done on the remaining *ESLO1* subset, on the interviews from *ESLO2,* and on the *forum* corpus. Two kinds of evaluation are performed:

- 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 (Cohen, 1960; Landis & Koch, 1977) measure;
- precision of automatic detection of the reformulation occurrences of the markers is evaluated against the manually created reference data. Precision is computed as ratio between the correct answers, found in the system results and the reference data, and the whole set of answers provided by the automatic system.

5. Results and their Discussion

Two main aspects are presented and discussed: inter-annotator agreement values and precision of the automatic system proposed for the distinction between reformulation occurrences of the studied markers (*c'est-à-dire, disons, ça veut dire*).

Corpus	*Agreement*	*Interpretation*
ESLO1	0.617	Substantial
ESLO2	0.526	Moderate
Forum	0.784	Substantial

Table 2. Inter-annotator agreement when deciding on presence of reformulations.

In Table 2, we indicate the inter-annotator agreement for the three corpora processed. On the whole, a substantial agreement is obtained on two corpora: *forum* and *ESLO1*. The agreement on the *ESLO2* data is moderate and is more difficult to obtain, which may be due to the conditions of collection of these data: the interviews are more informal and free by comparison with the *ESLO1* corpus. This fact may contribute to utterances which are longer, more heterogeneous and more complex. We can also observe that the best agreement is obtained on the written corpus *forum*. One reason for this may be that the written corpus, even if issued from online discussions, presents the final result of an idea, while spoken language keeps the elaboration of such idea (Hagège, 1985; Blanche-Benveniste et al., 1991; Blanche-Benveniste, 1995). For this reason, the discourse organization and the syntax of written corpora are more standardized and easier to process by the automatic tools.

Corpus	% of reformulations
ESLO1	26
ESLO2	35
Forum	61

Table 3. Rate of the reformulative occurrences of markers in the reference data.

In Table 3, we indicate the rate of reformulative occurrences of markers in the studied corpora. It reaches up to 61% in forum corpus, and 26% and 35% in *ESLO1* and *ESLO2*, respectively.

Set of rules	*ESLO1*	*ESLO2*	*Forum*
lexical and discursive filters	40.5	37.7	38.9

lexical and discursive filters + frequency (>6000)	25.8	18.8	40.3
lexical and discursive filters + priority frequency (>6000)	63.0	66.4	75.2

Table 4. Precision of the automatic detection of reformulations.

Precision of the rule-based system for the automatic detection of reformulations is indicated in Table 4. Like for the inter-annotator agreement, the rules-based system is more successful when processing the written corpus. When the lexical and discursive filters are applied alone, they reach up to 40%, 38% and 49% precision in *ESLO1, ESLO2* and *forum* corpora, respectively. The additional use of the frequency filters decreases substantially performance of the system. But, when the frequency filters have priority on the lexical and discursive filters, the overall precision is improved to up to 63%, 66% and 75%: in this case, we consider that frequency is indicative of the reformulation even if the utterance contains spoken language disfluencies. Notice that precision is improved 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 is higher than the inter-annotator agreement. It is comparable or even superior to the precision obtained in previous work (Bouamor, et al., 2012). By comparison with the paraphrase recognition results obtained on another written corpus, that is annotated mainly with lexical, syntactic and contextual paraphrases (Bhagat, 2013), our results are similar to those provided by the baselines and some of the systems reported (Androutsopoulos & Malakasiotis, 2010).

We expect to improve our current results in the next future.

6. Conclusion and Future Work

We propose a method for the automatic detection of reformulations in monolingual spoken *(ESLO1* and *ESLO2)* and written *(forum)* corpora in French. One originality is that we take into account the specificity of the spoken and written data through the building of utterances, the consideration of oral disfluencies, and the use of the NLP tool adapted to spoken (Dupont et al., 2012) and written (Laurent et al., 2009) corpora. Another originality is that we address the detection of reformulations with syntagmatic approach, within the *S1 marker S2* structure.

We perform manual and consensual annotation in order to obtain the reference data. Then, an automatic rule-based system is designed and tested for the detection of reformulative occurrences of the markers. The reference data allow evaluating the automatic method.

The best inter-annotator agreement is 0.784 and is observed on written corpus *forum.* On the spoken *ESLO1* and *ESLO2* corpora, the agreement is 0.617 and 0.526, respectively.

The automatic recognition of reformulative occurrences of markers relies on a set of filters (lexical, discursive and frequency) and reaches up to 75%. The comparison with the existing work confirms some previous observations (Rossari, 1990; Rossari, 1992): the markers

studied (*c'est-à-dire, disons, ça veut dire*) do not always introduce reformulations, and can perform other functions (e.g. argumentation, disfluency).

We have several directions for future work:

- we plan to involve additional annotators to obtain more annotated and consensual reference data;
- other markers can be studied and compared among them;
- for the automatic detection of reformulations, we can improve the current performance thanks to a better recognition of repetitions and to a machine learning approach;
- the automatic detection of boundaries of source *S1* and target *S2* entities in another perspective;
- spoken and written corpora will be further compared and analyzed from the point of view of reformulation: we assume the process is similar, as it allows making ideas clearer, and dissimilar from the cognitive point of view (Levelt, 1983; Hagege, 1985; Blanche-Benveniste et al., 1991);
- a similar study can be done on corpora from other languages;
- the annotations available further to this work will be made available to the research community. Besides, these annotations can also be stored in a TEI-compliant format.

Acknowledgement

We are thankful to the anonymous reviewers for the thorough reviewing and for their comments and suggestions which permitted to improve the quality and clarity of this article.

References

Androutsopoulos I. and Malakasiotis P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, **38,** 135-187.

Beeching K. (2007). 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)**,** 78-3.

Bhagat R. and Hovy E. (2013). What is a paraphrase? *Computational Linguistics*, **39**(3), 463-472.

Blanche-Benveniste C. (1995). Le semblamble et le dissemblable en syntaxe. *Recherches sur le français parlé*}, **13**, 7-33.

Blanche-Benveniste C., Bilger M., Rouget C. and Van Den Eynde K. (1991). *Le français parlé. Etudes grammaticales*. Paris: CNRS Editions.

Bouamor H., Max, A. and Vilnat, A. (2012). Etude bilingue de l'acquisition et de la validation automatiques de paraphrases sous-phrastiques. *TAL* **53**(1) (2012), 11-37.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37-46.

Culioli A. (1976). Notes du séminaire de DEA, 1983-84. Paris.

Dupont, Y., Tellier, I. and Courmet, A. (2012). *Un segmenteur-etiqueteur et un chunker pour le français*. Technical report, LIFO, Université d'Orléans (2012) demo.

Eshkol-Taravella I., Baude O., Maurel D., Hriba L., Dugua C. and Tellier I. (2012). Un grand corpus oral disponible : le corpus d'Orléans 1968-2012. *Traitement Automatique des Langues*, **52**(3), 17-46.

Eshkol-Taravella I. and Grabar N. (2014). Repérage et analyse de la reformulation paraphrastique dans les corpus oraux. In *TALN* 2014 .

Fløttum K. (1995). *Dire et redire. La reformulation introduite par "c'est-à-dire"*. Thèse de doctorat, Hogskolen i Stavanger, Stavanger.

Fuchs C. (1994). *Paraphrase et énonciation*. Paris: Orphys.

Fujita A. (2010). Typology of paraphrases and approaches to compute them. In *CBA to Paraphrasing & Nominalization*, Barcelona, Spain. Invited talk.

Gülich E. and Kotschi T. (1987). Les actes de reformulation dans la consultation. La dame de Caluire. In P. Bange, Ed., *L'analyse des interactions verbales. La dame de Caluire: une consultation*, p. 15-81. Berne: P Lang.

Hagège C. (1985). *L'homme de paroles. Contribution linguistique aux sciences humaines*. Fayard, Paris

Hölker, K. (1988). *Zur Analyse von Markern*. Franz Steiner, Stuttgart

Hwang, Y. (1993). Eh bien, alors, enn et disons en francais parle contemporain. *L'Information Grammaticale*, **57**, 46-48

Ibrahim A., Katz B. and Lin J. (2003). Extracting structural paraphrases from aligned monolingual corpora. In *International Workshop on Paraphrasing*, p. 57–64.

Kanaan L. (2011). *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.

Kok S. and Brockett C. (2010). Hitting the right paraphrases in good time. In *NAACL,* p. 145–153.

Landis J. and Koch G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33,** 159–174.

Laurent D., Nègre S. and Séguéla P. (2009). Apport des cooccurrences à la correction et à l'analyse syntaxique. In *TALN 2009*.

Le Bot M-C., Schuwer M. and Richard E. (dir.) (2008). La reformulation : Marqueurs linguistiques – Stratégies énonciatives. Rennes: Rivages linguistiques.

Levelt W. (1983). Monitoring and self-repair in speech. *Cognition,* **14,** 41-104.

Lin D. and Pantel L. (2001). Dirt - discovery of inference rules from text. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, p. 323–328.

Malakasiotis P. and Androutsopoulos I. (2007). Learning textual entailment using SVMs and string similarity measures. In *ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, p. 42-47.

Martin R. (1976). *Inférence, antonymie et paraphrase*. Paris: Klincksieck.

Melčuk I. (1988). Paraphrase et lexique dans la théorie linguistique sens-texte. In Lexique et paraphrase. *Lexique,* **6**, 13-54.

Neveu F. (2004). *Dictionnaire des sciences du langage*. Paris: Colin.

Petit, M. (2009). *Discrimination prosodique et representation du lexique : application aux emplois des connecteurs discursifs*. Thèse de doctorat, Université d'Orléans, Orléans

Quemada B. (1992). *FRANTEXT. Autour d'une base de données textuelles*. Publications de l'Institut National de la Langue Française, Didier Erudition

Rossari C. (1990). Projet pour une typologie des opérations de reformulation. *Cahiers de linguistique française*, **11,** 345-359.

Rossari C. (1992). De l'exploitation de quelques connecteurs reformulatifs dans la gestion des articulations discursives. Pratiques, 75, 111-124.

Roulet E. (1987). Complétude interactive et connecteurs reformulatifs. *Cahiers de linguistique française*, **8**, 111-140.

Sekine S. (2005). Automatic paraphrase discovery based on context and keywords between NE pairs. In *International Workshop on Paraphrasing*, p. 80–87.

Saunier E. (2012). Disons : un impératif de dire ? Remarques sur les propriétés du marqueur et son comportement dans les reformulations. *L'Information Grammaticale*, **132** , 25–34.

Teston-Bonnard, S. (2008). 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. Strategies enonciatives*. PUR, Rennes, 51-69

Vezin L. (1976). Les paraphrases: étude sémantique, leur rôle dans l'apprentissage. *L'année psychologique*, **76**(1), 177-197.

Vila M., Antonia Mart M. and Rodriguez H. (2011). Paraphrase concept and typology. A linguistically based and computationally oriented approach. *Procesamiento del Lenguaje Natural*, **46,** 83-90.

Vion R. (2006). Reprise et modes d'implication énonciative. *La Linguistique*, **2**(42), 11-28.