Why do we reformulate? Automatic Prediction of Pragmatic Functions

Natalia Grabar, Iris Eshkol

Why do we reformulate? Automatic Prediction of Pragmatic Functions

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

[1] CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France
natalia.grabar@univ-lille3.fr,
WWW home page: http://natalia.grabar.free.fr
[2] CNRS UMR 7270 LLL, Université d'Orléans, 45100 Orléans, France
iris.eshkol@univ-orleans.fr,
WWW home page: http://www.lll.cnrs.fr/iris-eshkol-taravella

Abstract. Reformulations participate in structuring of discourse, especially in dialogues, and also contribute to the dynamics of the discourse. Reformulation is a significant act which has to satisfy precise objectives. The purpose of our work is to automatically predict the reason for which a speaker performs a reformulation. We use a classification with eleven pragmatic functions inspired by the existing work and by the data analyzed. The reference data are built through manual and consensual annotations of spontaneous reformulations introduced by three markers (*c'est-à-dire, je veux dire, disons*) in French. The data are provided by spoken corpora and a corpus with forum discussions on health issues. We exploit supervised categorization algorithms and a set with several features (syntactic, formal, semantic and discursive) for the prediction of the reformulation categories. The distribution of sentences is not homogeneous across categories. The experiments are positioned at two levels: general and specific. Our results indicate that it is easier to predict the types of functions at the general level (the average F-measure is around 0.80), than at the level of individual categories (the average F-measure is around 0.40). We study the influence of various parameters.

Keywords: Reformulation, Machine Learning, Paraphrase, Classification, Pragmatic Function

1 Introduction

The notion of reformulation is central to our work and we will start with its definition and description. Reformulation consists of saying again and with different words an utterance or an idea. Usually, reformulation is done at the request of the interlocutor or by decision of the speaker.

The objective of our work is to study the reasons which make the speakers to reformulate. This is what we call the *pragmatic function of reformulations*. Two sources of reformulations are studied (spoken and forum corpora). Hence, we propose to analyze reformulated segments and to predict automatically the pragmatic function associated with each reformulation. The hypothesis of our work

is that the content of the reformulated segments $S1$ and $S2$ provides clues, be they non-linguistic (*e.g.*, size of the segments) or linguistic (*e.g.*, lexical, syntactic, semantic), for the prediction of these pragmatic functions. The experiments are performed at two levels: (1) at the general level, according to the type of linguistic transformations associated with reformulations: addition, reduction or equal amount of information; (2) at the specific level, through the exploitation of pragmatic functions (*e.g.*, definition, explanation, result, precision), such as described in section 2.4. We work with spontaneous reformulations in French, such as they occur in spoken language and in forum discussions. Reformulations are to be introduced by one of three markers studied (*c'est-à-dire, je veux dire, disons*) coined on the verb *dire* (*to say*). Besides, specific syntactic structure is studied: $S1$ *marker* $S2$.

In what follows, we first outline the notion of reformulation through the related work (section 2). We then describe the data we use (section 3) and the methodology we design (section 4). We present and discuss the results (section 5), and then conclude with some directions for future work (section 6).

2 Related work

2.1 Linguistic Work Applied to Written Corpora

In written language, reformulation may be related to several notions:

- *Paraphrase.* Reformulation can be seen as paraphrastic variation of linguistic segment in which formal modifications occur [38]. Hence, paraphrase is the result of reformulation. Notice that paraphrase is studied from different points of view: its relation to enunciation [15, 20, 21, 35, 44]; its linguistic transformations [36, 45, 7]; and its size [20, 22, 9].
- *Gloss.* Gloss is closely related to philological tradition and means commentaries done on a given word. It is composed of two parts: the first part is a lexical unit, while the second part is the gloss itself, often written in formal language [2, 43].
- *Repetition.* Repetition corresponds to various situations in which a given textual segment is repeated with various degrees of similarity [46]. The semantic proximity between the reformulated segments is then important.
- *Description.* Description is closely related with literary studies [32].
- *Elaboration.* Elaboration (of an idea) is a rhetoric relation which may also be similar to reformulation [39].

2.2 Linguistic Work Applied to Spoken Corpora

One of the main differences between spoken and written languages is that in spoken language we can observe the elaboration of ideas, while in written language the final result of this elaboration is usually presented [8]. Indeed, written language proposes the final version of the discourse and spoken language provides its creation with the hesitations, false starts, mistakes and reformulations, which

are all specific to it. Several existing works address spoken reformulation, which is one of its fundamental characteristics. Several objectives are addressed, although reformulation is always associated with disfluencies and self-corrections:

- Two types of reformulations may be distinguished [23, 41]: paraphrastic reformulation, which brings equivalence between the segments, and non-paraphrastic reformulation, which brings a change in enunciation perspective [40]. As each reformulation in spoken language does not provide paraphrase, two types of markers can distinguished: paraphrastic reformulation markers, like *c'est-à-dire, je veux dire, en d'autres termes*, and non-paraphrastic reformulation markers, like *en somme, de toute façon, enfin*.
- In syntactic studies of spoken language, reformulation is associated with enumeration or repetition [28, 8, 6], because repeated, reformulated or enumerated elements show the same syntactic place on the paradigmatic axis.
- Reformulation can also be associated with correction or precision [25].

2.3 Reformation and NLP

In the NLP field, reformulation in written corpora is often associated with paraphrase, as its result. Research questions are concerned with the automatic detection of paraphrases [4, 42, 3, 33, 30] and with their use in various applications, such as detection of plagiarism [19], textual entailment [1, 16], normalization of controlled languages [37], information retrieval and machine translation [31, 9].

In spoken corpora, reformulation is close to disfluencies and the existing works propose to detect it automatically. The methods used exploit manually crafted rules and patterns [10, 14] or supervised learning [17]. Detection of reformulations is then used for repairing, cleaning and reconstitution of enunciations.

2.4 Pragmatic Functions of Reformulation

Reformulation is a significant act performed with precise objective. This is what we call pragmatic function of reformulation, *i.e.* the role of spontaneous reformulation which can be observed in language. Reformulation links two segments: the source segments $S1$, which is reformulated, and the target segment $S2$, which contains the reformulation. As noticed above, we study reformulations formed with specific markers, derived from the verb *dire* (*to say*) (*c'est-à-dire, je veux dire, disons*), within the structure *S1 marker S2*.

We distinguish several pragmatic functions between $S1$ and $S2$, inspired from the existing typologies [23, 24, 5, 26] and motivated by our data:

- *Definition*: (often technical) terms from $S1$ are defined by $S2$: *avec une ETO c'est à dire une echographie tansoesophagienne (une écho ou le palpeur est introduit dans l'estomac)* (*by ETO I mean tansoesophagian echography (an echo in which a sensor is introduced in the stomach)*)
- *Explanation*: the speaker explains things to his interlocutor ($S2$ explains $S1$): *ce garçon je sais bien qu'il ne peut pas se marier avec euh c'est-à-dire qu'il*

aurait pu av- trouver une jeune fille euh qui fasse sa licence euh dans un milieu comme le nôtre (*this boy well I know that he cannot mary her euh I mean that he could find a girl euh who is doing studies euh from the same social standing*)
- *Exemplification*: for an entity mentioned in $S1$, examples are given in $S2$: *des morceaux nobles ce qu'ils appellent quoi c'est à dire les rosbifs les biftecks et tout ça* (*noble pieces what they call well I mean roast beef beefsteaks and all that*)
- *Justification*: $S2$ proposes justification of $S1$: *la langue française est plus difficile disons on peut pas dire la plus difficile des langues européennes mais c'est difficile* (*French is more difficult let's say one cannot say the most difficult of the European languages but it is difficult*)
- *Precision*: $S2$ is used to make clearer the statements from $S1$: *La trinitrine m'a été prescrite vendredi dernier, c'est à dire depuis une semaine* (*Trinitrin has been prescribed to me last Friday, I mean one week ago*)
- *Denomination*: $S2$ gives a name to an entity from $S1$: *en particulier c'est l'endroit où en somme ça s'est produit le plus au début c'est-à-dire à Nanterre* (*particularly this is the place where on the whole this happened at the beginning I mean in Nanterre*)
- *Result*: the speaker summarizes or indicates the consequence of $S1$: *A ma sortie, j'ai retrouvé pratiquement l'usage de ma jambe, de mon bras et ma main gauche, disons que je pouvais être autonome* (*At discharge, I almost recovered the usage of my left leg, arm and hand, let's say that I could be autonomous*)
- *Linguistic correction*: $S2$ makes linguistic correction (article, tense...) of $S1$: *des artisans euh hm hm hm hm hm alors c'est-à-dire artisans* (*the craftsmen euh hm hm hm hm hm so I mean craftsmen*)
- *Reference correction*: $S2$ corrects the place, time, etc. indicated in $E1$: *j'habitais rue Lazare Carnot c'est à dire donc au sud de la Source* (*I was living Lazare Carnot street I mean well on south from la Source*)
- *Paraphrase*: $S1$ repeats information from $S2$, but with different wording: *Il n'a acune maladie (je veux dire qu'il ne prend aucun médicament)* (*He has no disease (I want to say that he is not taking medication)*)
- *Opposition*: $S1$ repeats information from $S2$ in negative form: *il est joyeux je veux dire il n'est pas triste* (*he is happy I mean he is not sad*)

We can observe that occurrences of reformulation markers go far beyond their paraphrastic usage and cover a large set of situations.

3 Processed and Exploited Data

We work with several types of data: (1) two types of corpora (two ESLO corpora (section 3.1) and corpus with discussions from medical forum (section 3.2)), (2) reformulated segments (section 3.3) gained with a manual and consensual annotation of corpora, and (3) several linguistic resources (section 3.4).

3.1 ESLO

ESLO (Enquêtes Sociolinguistiques à Orléans), *ESLO1* and *ESLO2*, corpora [18] are corpora with spoken data in French. *ESLO1* has been collected in 1968-1971

Corpus	*Agreement*	*Interpretation*
ESLO1	0.617	substantial
ESLO2	0.526	moderate
Forum	0.784	substantial

Table 1. Inter-annotator agreement on the presence of reformulations in sentences.

by French teaching staff from Essex university and members of the B.E.L.C. (Bureau pour l'étude de l'enseignement de la langue et de la civilisation françaises de Paris). This corpus contains 300 hours of speech (4,500,000 words). Collection of *ESLO2* started in 2008. This corpus will contain over 350 hours of speech (5,500,000 words). These two corpora are available online[3].

3.2 Discussion Forum

Forum corpus has been collected from online discussions in *Hypertension* forum[4]. It provides 12,588 threads with 67,652 posts and 6,788,361 word occurrences. The posts are written by users, whose need is to speak about their medical problems. These are non-normed writings with frequent misspellings and errors, and specific non conventional linguistic elements (abbreviations, emoticons...).

3.3 Reformulated Segments

The 4,120 sentences with three reformulation markers, extracted from our corpora (*ESLO1*, *ESLO2*, *forum*), have been manually annotated by two independent annotators and have gone through consensus. The inter-annotator agreement is computed with the Cohen kappa [13]. In Table 1, we indicate the obtained agreement, computed on the judgment on the presence of reformulations, and its standard interpretation [27]. We can see that the agreement is moderate and substantial. When the agreement is computed at the level of pragmatic functions, it is very low: 0.127 on *ESLO1* and 0.0211 on *ESLO2*.

Among the 4,120 marker occurrences, 594 bring reformulations. In Table 2, we indicate their distribution across corpora and pragmatic functions. We can see that *precision* is the most used function. *Linguistic correction* and *opposition* are very rare, and their relevance for our work can be reconsidered. *Justification* has similar distribution in *ESLO2* and *forum*, and *paraphrase* in the three corpora. Other functions are not evenly distributed across the corpora. For instance, *definition* is very frequent in forum discussions (medical terms and their definitions are important there), *exemplification* and *explanation* are frequent in *ESLO1* and *ESLO2*. *Denomination* is rare in spoken, but frequent in *forum* corpora. For instance, in forum discussions, it allows to name medications and treatments. These observations indicate that the nature of corpora and of pragmatic functions impact their real use by speakers.

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

[4] http://forum.doctissimo.fr/sante/hypertension-problemes-cardiaques/liste_sujet-1.htm

Function	*ESLO1*	*ESLO2*	*ESLO*	*forum*	*total*
cor-ling	-	2	2	-	2
cor-ref	5	1	6	-	6
def	16	14	30	41	71
denom	2	3	5	24	29
exempl	29	15	44	21	65
explic	26	16	42	25	67
justif	1	8	9	8	17
oppo	2	-	2	-	2
para	14	18	32	20	52
prec	47	54	101	88	189
res	19	43	62	32	94
total	161	174	335	259	594

Table 2. Distribution of sentences between the pragmatic functions and corpora.

3.4 Linguistic Resources

We exploit several types of resources: (1) list with stopwords; (2) disfluency markers; (3) distributional clusters with words generated from our corpora; (4) lexicon with hyponyms; (5) specific lexical markers.

Stopwords. Stopwords (n=69) are mainly grammatical words. They are used to remove non-relevant content and to make the statistical processing more rapid.

Disfluency markers. We use a set of disfluency markers: *allez, allons, alors, là, enfin, euh, heu, bah, ben, hm, hum, hein, quoi, ah, oh, donc, bon, bè, eh.*

Clusters with words. Distributional clusters with words are generated from our corpora: *ESLO1*, *ESLO2*, *ESLO* (merging of *ESLO1* and *ESLO2*), *forum* and all the corpora together (*total*). The corpora are segmented and lower-cased, the stopwords are removed. The clusters are generated with the existing clustering algorithm [12, 29]. This is hierarchical agglomerative clustering based on distributional information on words. Within a given cluster, the words are semantically related because they occur in similar contexts. We generate distributional resources with 200 to 600 clusters.

Hyponyms. A lexicon of hyponyms is automatically extracted from Wiktionary[5] in French. The structure of Wiktionary articles is exploited for the extraction of entries and their hyperonyms. This lexicon contains 12,161 pairs {*hyperonym*; *hyponym*}, such as {*lexique*; *dictionnaire*} (*{lexicon; dictionary}*), {*armée*; *légion*} (*{army; legion}*), {*disque*; *CDROM*} (*{disk; CDROM}*), {*période*; *année*} (*{period; year}*). Words from a given pair have strong semantic link.

Lexical markers. We use a small set of markers (n=17) associated with pragmatic functions. Three types of markers are distinguished: (1) introductory markers (*e.g., voilà (well), c'est (is a), ce sont (is a in plural)*), which can occur with definitions; (2) causal markers (*e.g. c'est pourquoi (this is why), parce que (because), car (because)*), which can occur with *result*; (3) exemplification

[5] https://fr.wiktionary.org/wiki/Wiktionnaire

markers (*e.g. exemple (example), comme (such as), entre autre (among other)*), which can occur with the *exemplification* function.

4 Methods

The main steps of the method are: (1) pre-processing and creation of the reference data; (2) supervised categorization of segments in order to predict their pragmatic function; and (3) evaluation. We perform supervised categorization through the Weka platform [34] and use several algorithms available in their standard configuration. We describe now the reference data, the categories and the features used, as well as the evaluation modalities.

4.1 Reference data

The reference data are obtained from the manual and consensual annotations of sentences. Table 2 presents these data, according to the pragmatic functions and corpora. Two types of corpora are processed: spoken corpora *ESLO* and corpus *forum* with forum discussions. We can see that reformulations are distributed homogeneously between these two types of corpora, while some functions are over-represented (*precision*, *result*, *definition*, *explanation* and *exemplification*).

4.2 Categories

Categories correspond to pragmatic functions from Table 2. Because three categories (*linguistic correction, referential correction, opposition*) are very small, we perform experiments with eight most frequent categories. As noticed, another experiment is positioned at a more general level, according to the amount of information provided during the reformulation and measured by the size of segments:

- reduction of information in $S2$ by comparison with $S1$: *result, denomination*;
- addition of information in $S2$ by comparison with $S1$: *definition, exemplification, explanation, justification, precision*;
- equal amount of information: *paraphrase, linguistic correction, referential correction, opposition.*

This typology is similar to the one proposed in existing work in literary studies [32], although we also distinguish the reduction of information in $S2$. This rationale permits to perform experiments at two levels: general level with three categories (addition, reduction and comparable amount of information), and specific level with eight categories.

4.3 Features

We exploit several features in order to describe the nature of pragmatic functions. The values of these features are transformed in numerical values:

- length of segments $S1$ and $S2$, in words and characters,
- difference of length of segments $S1$ and $S2$, in words and characters,
- equivalence between syntactic categories of the two segments,
- whether the syntactic category of segments is nominal group or proposition,
- presence of segments, or of their words, in the same clusters. Several versions are used: all words, all words without identical words, all words without stopwords, all words without identical and stopwords. Numbers and rates of common words are computed. We use several sets of clusters according to the corpora they are generated from (*ESLO1*, *ESLO2*, *ESLO*, *forum* and all corpora together (*total*)), and to the number of clusters to be generated (we exploit 300 and 600 clusters in the analysis of the results),
- presence of disfluency markers in segments,
- presence of numbers in segments,
- presence of upper-cased characters in segments,
- presence of specific lexical markers (exemplification, causal markers and introductory structures),
- presence of segments, or of their words, in pairs of words with hyperonymy relation.

As we can see, the features are positioned at different levels: formal, syntactic, semantic and discursive. These features are computed automatically, with or without exploitation of linguistic resources.

4.4 Evaluation

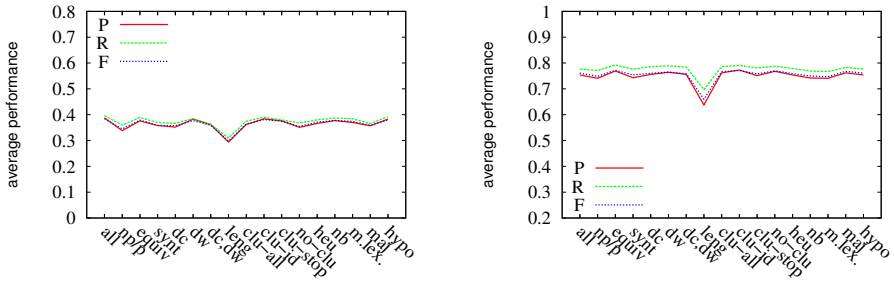
Evaluation is performed with classical evaluation measures: precision, recall and F-measure. We present the evaluation figures such as computed by the Weka platform. We perform a 10-fold cross-validation: the data are segmented in 10 folds and, at each iteration, one fold is used for training while the rest of folds is used for the testing. The global evaluation corresponds to the average of evaluations obtained at each iteration.

5 Results

In preliminary experiments, we observed that `RandomForest` [11] optimizes prediction of categories. With the data taken all together, `RandomForest` provides the average 0.38 with eight categories, and 0.76 with three categories. We present the results obtained with this algorithm.

Different parameters and features impact the results:

- in Figure 1, we indicate the average performance (precision, recall, F-measure) obtained when various features are removed from the feature set. On the whole, performance remains close to when all the features are used (experience with all features *all* is at the first position): 0.4 with eight categories and 0.8 with three categories. Notice that removal of some features (equivalence of syntactic categories *equiv*, information on clusters (*clu-all, clu-in,*



(a) Removed features, 8 categories

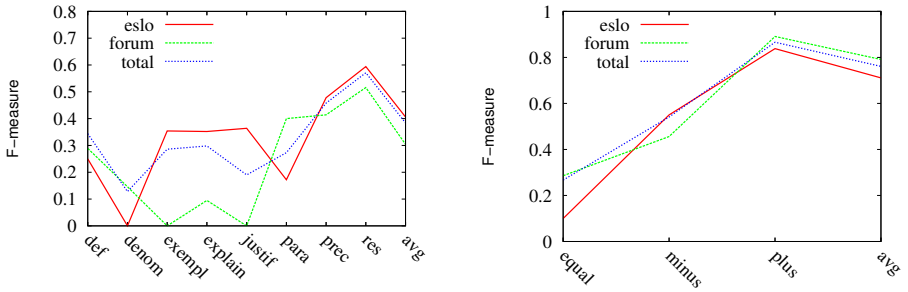
(b) Removed features, 3 categories

Fig. 1. Removal of different features at each experiment.

clu-stop, no-clu), upper-cased characters (*maj*)) is benefic for the prediction of three categories (Figure 1(b)), while with eight categories the whole set of features is always more efficient than when some of features are removed (Figure 1(a)). Removal of information on the size of segments $S1$ and $S2$ *leng* always causes an important decrease of performance;

- with the whole set of features, the most efficient feature is the size difference in characters between $S1$ and $S2$. With this feature alone, the global F-measure is 0.28 and 0.73, with eight and three categories, respectively. Other features related to the length of $S1$ and $S2$ are also important. When these features are removed, presence of disfluency markers becomes the most important feature;
- distributional resource has no difference between *total* and *forum* corpora. Yet, with *ESLO2*, it is suitable to use distributional resources generated from the same corpus or from the two *ESLO* corpora: the nature and the content of spoken corpora clearly present linguistic specificities;
- Figure 2 indicates the recognition of pragmatic functions in three corpora (*ESLO*, *forum* and *total*). With eight categories, we observe that *result* function is the best recognized in all the corpora. *precision* and *definition* show less efficient prediction, although it is stable across corpora. *paraphrase* is quite well recognized in *forum*, but poorly in other corpora, while *exemplification*, *explanation* and *justification* are quite well recognized in *ESLO*. With three categories, the *plus* category is the easiest to recognize. These observations must be correlated with the amount of the reference data in each corpus and category: *result*, *precision* and, by consequence the *plus* categories, are the biggest categories in our data set.

An analysis of the confusion matrix with eight categories indicates that some functions are very close and often confounded. For instance, the *precision* function is often confounded with other functions. A possible explanation comes from the nature of this function. *Precision* seems to be a large category which can contain *explanation*, *definition*, *exemplification*, *denomination*. It is necessary to



(a) 8 categories (b) 3 categories

Fig. 2. Performance in recognition of pragmatic functions in each corpus.

define additional and formal constraints to make this function more specific. Another reason is that this function is very frequent, by comparison with other functions, which can also favor its automatic recognition. We can also notice that the *denomination* category brings numerous confusions: almost all its instances are categorized in other categories.

6 Conclusion and Future Work

We propose to study reformulation in two types of corpora: spoken and written (forum discussions). We concentrate on pragmatic functions of reformulations, which correspond to the reason of why speakers perform a given reformulation. We have built a classification with eleven functions (*e.g. definition, exemplification, result, paraphrase, linguistic correction*). The purpose of our work is to study these functions and to predict them, thanks to the analysis of the content of segments $S1$ and $S2$ related by three reformulation markers (*c'est-à-dire, je veux dire, disons*). Exploitation of consensual reference data and of supervised machine learning algorithms permits to do experiments at two levels: (1) at the general level, which categories are related to the amount of information in $S2$ (addition, reduction or equal amount of information), we obtain performance close to 0.80; (2) at the specific level with eight individual categories, we obtain performance close to 0.40. Some features (related to the length of segments and to disfluencies) play an important role for the prediction of pragmatic functions.

Prediction and discovery of information of the pragmatic nature is extremely difficult in language. Hence, we can consider that our work is rather exploratory and permits to find out various points which must be taken into account in future work. From the linguistic point of view, it will be important to reconsider some functions: *linguistic correction*, *opposition*, and *precision*. The last function should be described with additional formal criteria. From the computational point of view, additional features and other algorithms should be used for improving the prediction of pragmatic functions of reformulations.

References

1. Androutsopoulos, I., Malakasiotis, P.: A survey of paraphrasing and textual entailment methods. Journal of Artificial Intelligence Research 38, 135–187 (2010)
2. Authier-Revuz, J.: Ces mots qui ne vont pas de soi : boucles réflexives et non-coïncidences du dire. Larousse, Paris (1995)
3. Bannard, C., Callison-Burch, C.: Paraphrasing with bilingual parallel corpora. In: ACL. pp. 597–604 (2005)
4. Barzilay, R., McKeown, L.: Extracting paraphrases from a parallel corpus. In: ACL. pp. 50–57 (2001)
5. 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), 78–93 (2007)
6. Benzitoun, C.: L'annotation syntaxique de corpus oraux constitue-t-elle un problème spécifique ? In: RECITAL 2004 (2004)
7. Bhagat, R., Hovy, E.: What is a paraphrase? Computational Linguistics 39(3), 463–472 (2013)
8. Blanche-Benveniste, C., Bilger, M., Rouget, C., Van Den Eynde, K.: Le français parlé. Études grammaticales. CNRS Éditions, Paris (1991)
9. Bouamor, H.: Étude de la paraphrase sous-phrastique en traitement automatique des langues. Thèse de doctorat, Université Paris Sud, Paris (2012)
10. Bouraoui, J.L., Vigouroux, N.: Analyse des erreurs de performance et des stratégies correctives dans le dialogue oral spontané : apports à l'étude des pathologies du langage. Revue Parole 29-30, 121–152 (2004)
11. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
12. Brown, P., deSouza, P., Mercer, R., Della Pietra, V., Lai, J.: Class-based n-gram models of natural language. Computational Linguistics 18(4), 467–479 (1992)
13. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1), 37–46 (1960)
14. Constant, M., Dister, A.: Automatic detection of disfluencies in speech transcriptions, pp. 259–272 (2010)
15. Culioli, A.: Notes du séminaire de DEA, 1983-84. Paris (1976)
16. Dagan, I., Roth, D., Sammons, M., Zanzotto, F.: Recognizing Textual Entailment. Morgan & Claypool Publishers, Milton Keynes, UK (2013)
17. Dutrey, C., Clavel, C., Rosset, S., Vasilescu, I., Adda-Decker, M.: A CRF-based approach to automatic disfluency detection in a French call-centre corpus. In: International Speech Communication Association Conference (INTERSPEECH 2014). p. 5 (2014)
18. 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 des Langues 52(3), 17–46 (2012)
19. Ferrero, J., Simac-Lejeune, A.: Détection automatique de reformulations – correspondance de concepts appliquée à la détection de plagiat. In: EGC 2015, RNTI-E-28. pp. 287–298 (2015)
20. Flottum, K.: Dire et redire. La reformulation introduite par "c'est-à-dire". Thèse de doctorat, Hogskolen i Stavanger, Stavanger (1995)
21. Fuchs, C.: Paraphrase et énonciation. Orphys, Paris (1994)
22. Fujita, A.: Typology of paraphrases and approaches to compute them. In: CBA to Paraphrasing & Nominalization. Barcelona, Spain (2010), invited talk

23. 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, pp. 15–81. P Lang, Berne (1987)
24. Hölker, K.: Zur Analyse von Markern. Franz Steiner, Stuttgart (1988)
25. Kahane, S., Pietrandrea, P.: La typologie des entassements en français. In: CMLF 2012. pp. 1809–1828 (2012)
26. 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)
27. Landis, J., Koch, G.: The measurement of observer agreement for categorical data. Biometrics 33, 159–174 (1977)
28. Levelt, W.: Monitoring and self-repair in speech. Cognition (14), 41–104 (1983)
29. Liang, P.: Semi-Supervised Learning for Natural Language. Master, Massachusetts Institute of Technology, Boston, USA (2005)
30. Lin, D., Pantel, L.: Dirt - discovery of inference rules from text. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 323–328 (2001)
31. Madnani, N., Dorr, B.J.: Generating phrasal and sentential paraphrases: A survey of data-driven methods. Computational Linguistics 36, 341–387 (2010)
32. Magri-Mourgues, V.: Reformulation et dialogisme dans le récit de voyage (2013)
33. Malakasiotis, P., Androutsopoulos, I.: Learning textual entailment using SVMs and string similarity measures. In: ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. pp. 42–47 (2007)
34. Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge, MA (1999)
35. Martin, R.: Inférence, antonymie et paraphrase. Klincksieck, Paris (1976)
36. Melčuk, I.: Paraphrase et lexique dans la théorie linguistique sens-texte. In Lexique et paraphrase. Lexique 6, 13–54 (1988)
37. Nasr, A.: Un modèle de reformulation automatique fondé sur la théorie sens-texte: application aux langues contrôlées. Thèse de doctorat, Université Paris 6 (1996)
38. Neveu, F.: Dictionnaire des sciences du langage. Colin, Paris (2004)
39. Péry-Woodley, M.P., Asher, N., Enjalbert, P., Benamara, F., Bras, M., Fabre, C., Ferrari, S., Ho-Dac, L.M., Le Draoulec, A., Mathet, Y., Muller, P., Prévot, L., Rebeyrolle, J., Tanguy, L., Vergez-Couret, M., Vieu, L., Widlöcher, A.: ANNODIS: une approche outillée de l'annotation de structures discursives. In: TALN 2009 (2009)
40. Rossari, C.: Projet pour une typologie des opérations de reformulation. Cahiers de linguistique française 11, 345–359 (1990)
41. Roulet, E.: Complétude interactive et connecteurs reformulatifs. Cahiers de linguistique française 8, 111–140 (1987)
42. Shinyama, Y., Sekine, S., Sudo, K., Grishman, R.: Automatic paraphrase acquisition from news articles. In: Proceedings of HLT. pp. 313–318 (2002)
43. Steuckardt, A.: Les marqueurs formés sur dire, pp. 51–65 (2005)
44. Vezin, L.: Les paraphrases: étude sémantique, leur rôle dans l'apprentissage. L'année psychologique 76(1), 177–197 (1976)
45. Vila, M., Antònia Mart, M., Rodríguez, H.: Paraphrase concept and typology. a linguistically based and computationally oriented approach. Procesamiento del Lenguaje Natural 46, 83–90 (2011)
46. Vion, R.: Reprise et modes d'implication énonciative. La Linguistique 2(42), 11–28 (2006)