Detection of Reformulations in Spoken French

Natalia Grabar, Iris Eshkol-Taravela

Detection of Reformulations in Spoken French

Natalia Grabar[1], **Iris Eshkol-Taravela**[2]
[1]CNRS, UMR 8163, F-59000 Lille, France;
Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
[2]CNRS UMR 7270 LLL, Université d'Orléans, 45100 Orléans, France,
natalia.grabar@univ-lille3.fr, iris.eshkol@univ-orleans.fr

Abstract

Our work addresses automatic detection of enunciations and segments with reformulations in French spoken corpora. The proposed approach is syntagmatic. It is based on reformulation markers and specificities of spoken language. The reference data are built manually and have gone through consensus. Automatic methods, based on rules and CRF machine learning, are proposed in order to detect the enunciations and segments that contain reformulations. With the CRF models, different features are exploited within a window of various sizes. Detection of enunciations with reformulations shows up to 0.66 precision. The tests performed for the detection of reformulated segments indicate that the task remains difficult. The best average performance values reach up to 0.65 F-measure, 0.75 precision, and 0.63 recall. We have several perspectives to this work for improving the detection of reformulated segments and for studying the data from other points of view.

Keywords: Spoken Corpora, Reformulation, Reformulation Marker, Paraphrase, Supervised Machine Learning

1. Introduction

Reformulations may occur in written and spoken languages, in which they show different functions (Flottum, 1995; Rossari, 1992): in spoken language, they mark the elaboration of ideas, and are punctuated by hesitations, false starts, and repetitions (Blanche-Benveniste et al., 1991), in written documents, we usually find the final result of the reformulation process (Hagège, 1985). It is 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 (Gülich and Kotschi, 1987; Kanaan, 2011). Specific reformulation markers may provide the formal mark-up of reformulations.

Reformulation is closely related to paraphrases, in that way that reformulated sequences can produce the paraphrases (Neveu, 2004). Reformulation and paraphrase play an important role in languages:

- When studying languages, a common exercise consists of paraphrasing expressions in order to control their understanding by students;
- In the same way, it is possible to control the understanding of ideas. The first exercises of the kind have appeared with the exegesis of ancient texts: sacred texts (Bible, Koran, Torah) first, and then theological, philosophical and scientific texts;
- More naturally, speakers use the reformulation and paraphrase in order to precise and to better transmit their thoughts. It is also common to find reformulations in written language: between various versions of the same literary piece of work (Fuchs, 1982), of the Wikipedia articles (Vila et al., 2014), or of scientific articles. The authors can thus rewrite several times their text until they produce the one that suits them at last.

Reformulation and paraphrase also play an important role in different NLP applications (Madnani and Dorr, 2010; Androutsopoulos and Malakasiotis, 2010; Bouamor et al., 2012). The objective is to detect linguistic expressions that differ by their form but convey the same or similar meaning:

- In information retrieval and extraction, paraphrases permit to increase the coverage of the found or extracted results. For instance, pairs like {*myocardial infarction*, *heart attack*} and {*Alzheimer's disease*, *neurodegenerative disease*} contain different expressions that convey identical or close semantics;
- In machine translation, paraphrases permit to avoid lexical repetitions (Scarpa, 2010);
- Textual entailment (Dagan et al., 2013) consists of creating relation between two textual segments, called Text and Hypothesis. Entailment is a directional relation, in which the truth of the Hypothesis must be inferred through the analysis of the Text. For instance, the Text *The drugs that slow down or halt Alzheimer's disease work best the earlier you administer them* allows inferring that the Hypothesis *Alzheimer's disease is treated by drugs* is true; while the Hypothesis *Alzheimer's disease is cured by drugs* cannot be inferred from this Text. In this example, the paraphrases {*administer drugs*, *treated by drugs*} permit to establish the right link between the Text and the Hypothesis.

As these few examples indicate, reformulation and paraphrase may cover various linguistic phenomena. The corresponding classifications may be more or less complex: from 25 (Bhagat and Hovy, 2013) to 67 (Melčuk, 1988) categories. Most often, these classifications address one given aspect, such as linguistic characteristics (Melčuk, 1988; Vila et al., 2011; Bhagat and Hovy, 2013), size of the paraphrased units (Flottum, 1995; Fujita, 2010; Bouamor, 2012), knowledge required for understanding the paraphrastic relation (Milicevic, 2007), language register. To our knowledge, there is only one multidimensional classification of paraphrase (Milicevic, 2007). In our work, we also propose to use a multidimensional classification, that covers the following dimensions, some of which are inspired by the previous works (Gulich and Kotschi, 1983; Beeching, 2007; Vila et al., 2011):

- syntactic category of the reformulated segments,
- type of lexical relation between the segments (*e.g.* hyperonymy, synonymy, antonymy, instance, meronymy),
- type of lexical modification (*e.g.* replacement, removal, insertion),
- type of morphological modification (*i.e.* inflection, derivation, compounding),
- type of syntactic modification (*e.g.* passive/active way),
- type of pragmatic relation between the reformulated segments (*e.g.* definition, explanation, precision, result, linguistic correction, referential correction, equivalence).

2. Existing works in the automatic acquisition of paraphrases

Several approaches have been proposed for the automatic detection of paraphrases. As explained above, in our work, we associate reformulation and paraphrase, which can be seen as the result of reformulation. Usually, the existing approaches exploit paradigmatic properties of words and their capacity to replace each other in a given context. These approaches depend on the corpora exploited. Four types of corpora are usually distinguished: monolingual, monolingual parallel, monolingual comparable, and bilingual parallel.

Monolingual corpora. Two kinds of approaches may be used with monolingual corpora:

- computing the similarity of strings permits to detect linguistic units (words, expressions, etc.) that show common surface features such as with {*When did Charle de Gaulle die?*, *Charles de Gaulle died in 1970*} (Malakasiotis and Androutsopoulos, 2007),
- distributional methods allow to detect units that occur in similar contexts. Such units have similar contextual or syntactic vectors, and may be good candidates for the paraphrase (*e.g.* {*Y is solved by X*, *Y is resolved in X*}) (Lin and Pantel, 2001; Paşca and Dienes, 2005).

Monolingual parallel corpora. When a given text is translated more than once in another language, these translations allow to build monolingual parallel corpora. One of the most used corpora is Jules Verne's *20 000 lieux sous la mer* that has been translated twice in English. Once these corpora are aligned at the sentence level, it is possible to exploit them with word alignment tools (Och and Ney, 2000). Various methods have been proposed for such exploitation (Ibrahim et al., 2003; Quirk et al., 2004; Barzilay and McKeown, 2001). They allow to extract paraphrases such as {*countless*, *lots of*}, {*undertone*, *low voice*}, {*shrubs*, *bushes*}, {*refuse*, *say no*}, {*dull tone*, *gloom*} (Barzilay and McKeown, 2001).

Monolingual comparable corpora. Monolingual comparable corpora typically contain texts on the same event but created independently, such as news articles. The thematic coherence of these texts and the distributional methods or alignment of comparable sentences may lead to the detection of paraphrases (Shinyama et al., 2002; Sekine, 2005; Shen et al., 2006). More particularly, named entities and numbers are part of the clues used for the extraction of paraphrases, such as in {*PERS1 killed PERS2*, *PERS1 let PERS2 die from loss of blood*} or {*PERS1 shadowed PERS2*, *PERS1 kept his eyes on PERS2*} (Shinyama et al., 2002).

Bilingual parallel corpora. Bilingual parallel corpora typically contain translations of a given text in another language. Once they are aligned at the level of sentences, they can also be used for the detection of paraphrases. Different translations of a given linguistic unit can provide paraphrases (Bannard and Callison-Burch, 2005; Callison-Burch et al., 2008; Kok and Brockett, 2010). For instance, the paraphrases {*under control*, *in check*} can be extracted because they are translations of *unter kontrolle* (Bannard and Callison-Burch, 2005).

3. Objectives

We have multi-fold objectives:

1. propose annotation guidelines for reformulations and to test them when annotating enunciations from French spoken corpora. These guidelines are presented in a previous work (Eshkol-Taravella and Grabar, 2014) and outlined at the end of Section 1. They allow creating the reference data;
2. study three specific reformulation markers: *c'est-à-dire* (*in other words*), *je veux dire* (*that is to say / I mean*), and *disons* (*let's say*). Several other reformulation markers exist (*notamment, en d'autres mots, en d'autres termes...*), but we propose to concentrate here on these three markers, coined on the verb *dire* (*to say*);
3. work with specific structure: *source-entity marker target-entity*, such as in example (1), in which the two entities *démocratiser l'enseignement (democratize the education)* and *permettre à tout le monde de rentrer en faculté (allow everybody to enter the university)* show the reformulation relation expressed syntagmatically;
4. distinguish enunciations that contain reformulations around the markers studied. This aspect is processed by the method proposed in Section 5.2.;
5. detect the segments that are reformulated: {*démocratiser l'enseignement*, *permettre à tout le monde de rentrer en faculté*} ({*democratize the education, allow everybody to enter the university*}). This aspect is processed by the method proposed in Section 5.3.

(1) *démocratiser l'enseignement c'est-à-dire permettre à tout le monde de rentrer en faculté* (*democratize the education in other words allow everybody to enter the university*) [ESLO1_ENT_121_C]

4. Linguistic Data Used

4.1. Reformulation Markers (RM)

We use three reformulation markers (RM): *c'est-à-dire* (Gulich and Kotschi, 1983; Hölker, 1988; Beeching, 2007), *je veux dire* (Teston-Bonnard, 2008) and *disons* (Hwang, 1993; Petit, 2009; Saunier, 2012). The common point between them is that they are coined on the basis of the verb *dire* (*to say*). In the existing work, these markers are recognized for their capacity to introduce reformulations, although they can play other roles, such as argumentation and disfluencies.

4.2. Corpora processed

We work with the ESLO (Enquêtes Sociolinguistiques à Orléans) corpora (Eshkol-Taravella et al., 2012): *ESLO1* and *ESLO2*. *ESLO1* has been created in 1968-1971 by French teachers from the Essex university, Language Centre, Colchester, UK, in collaboration with the B.E.L.C. lab (Bureau pour l'étude de l'enseignement de la langue et de la civilisation françaises de Paris). *ESLO1* contains 300 hours of speech (4,500,000 words approximately) and includes several types of recordings. Building of *ESLO2* started in 2008. It will contain over 350 hours of speech. *ESLO1* and *ESLO2* are accessible online (*http://eslo.tge-adonis.fr/*).

4.3. 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*.

4.4. Reference data

In the reference data, the reformulations are annotated through a multidimensional classification: syntactic, lexical, morphological and functional properties (see example (2)) annotated using dedicated guidelines.

(2) *euh <VP1>démocratiser l'enseignement </VP1> <RM>c'est-à-dire</RM> <VP2 rel-lex="syno(démocratiser/permettre à tout le monde) mero(enseignement/faculté)" modif-lex="ajout(rentrer à)" rel-pragm= "explic"> permettre à tout le monde de rentrer en faculté</VP2>* [ESLO1_ENT_121_C]
(*euh <VP1>democratize the studies</VP1> <RM>in other words</RM> <VP2 rel-lex="syno(democratize/allows everybody) mero(study/university)" modif-lex="insertion(enter)" rel-pragm="explic"> allow everybody to enter the university</VP2>*)

Three annotator pairs have participated in the creation of the reference data. After the independent annotation, consensus has been reached. The inter-annotator agreement is computed on the judgment of whether a given enunciation contains a reformulation or not (in which case, the marker introduces disfluency, argumentation...). The inter-annotator agreement (Cohen, 1960) is substantial 0.617 in *ESLO1* and moderate 0.526 in *ESLO2* (Landis and Koch, 1977). On the whole, 611 enunciations in *ESLO1* and 498 enunciations in *ESLO2* are annotated. Currently, 168 reformulations are provided by *ESLO1* and 186 by *ESLO2* (59 and 37 recordings, respectively). The rate of reformulations is 28% in *ESLO1* and 37% in *ESLO2*. These reference data are used for making observations, for training the system and for evaluating the automatically obtained results.

5. Approaches for the processing and detection of reformulations

In Figure 1, we present the general schema of the method composed of several steps: preprocessing of data (Section 5.1.), detection of enunciations with reformulations (Section 5.2.). Only those that contain reformulations are processed further. We then perform the detection of reformulated segments (Section 5.3.) and evaluate the results (Section 5.4.).

5.1. Preprocessing of data

Transcriptions from the ESLO corpora have adopted standard French spelling and non-use of punctuation. The original segmentation is done with the intuitive unit *breath group* detected by human transcribers, or with the *speech turn* detected with the change of speakers. We have rebuilt the enunciations using the speech turns with the change of speakers, and the overlappings when two or more speakers speak at the same time. With overlappings, the corresponding segments are associate to enunciations of each of the involved speakers. The processed unit corresponds to one enunciation.

Enunciations are processed with the SEM chunker (Tellier et al., 2014) adapted to French spoken language in order to detect the minimal chunks.

5.2. Detection of enunciations with reformulations

We consider that there is no reformulation if:

1. the context is not sufficient (beginning or end of enunciation);

2. markers occur with disfluencies, primes (*s-*), etc., (Blanche-Benveniste et al., 1991);

3. markers occur in specific contexts (use of *nous* (*we*) with *disons* then meaning (*we say*));

4. markers are found within existing expressions, like *indépendamment de* (*independently on*) (example (3)). The last test is done with the chunker output without markers and disfluencies: we record the frequencies obtained for the tested segments. Several thresholds of frequency are tested (*e.g.* 60, 80, 100). If the observed frequency is higher than the thresholds, we consider that the segment contains an existing expression with disfluency.

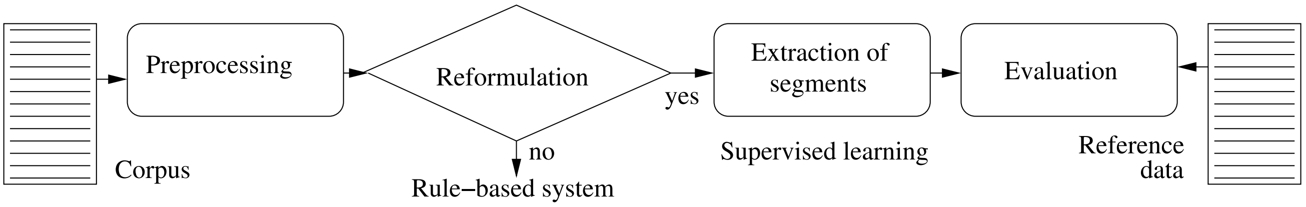

Figure 1: General schema of the method.

(3) *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 disons de leurs oui origines de classe* [ESLO1 _ENT_001_C]
(*do you see an important difference between your clients in their ways to choose the meat in what they buy et caetera independently say on their social origins*)

5.3. Extraction of reformulated segments

The boundaries of the reformulated segments are detected with a CRF machine learning algorithm (Lavergne et al., 2010). The reference data are divided in training and test sets with 60% and 40% of the enunciations, respectively.

5.3.1. Categories to be detected

The objective is to detect the two reformulated segments: the source segment, that is reformulated later in the text, and the target segment, that proposes new linguistic expression of the idea already expressed by the source segment. The categories to be detected are the following:

1. **M**: reformulation marker,
2. **SEG1**: source segment, which occurs before the marker,
3. **SEG2**: target segment, which occurs after the marker,
4. **O**: other tokens (out position).

5.3.2. Set of features

Each word is described with a set of features (Table 1):

- *form*: form of each word as it occurs in text;
- *POS*: POS tag computed by SEM;
- *chunksBI*: SEM chunks with beginning and end markup;
- *chunks*: SEM chunks without beginning and end mark-up;
- *heu*: we check out whether the word is part of the list with the disfluency markers (Section 4.3.);
- *num*: number of each word;
- *début/milieu/fin*: relative position of each word. Possible values: beginning (first 20% of words), end (last 20% of words), and middle MIL (all the rest);
- *stem*: words stemmed by `Snowball` (Porter, 1980);
- *RM*: mark-up of the RM.

5.3.3. CRF patterns

The CRF patterns provide the possibility to indicate how the features must be exploited: their combinations and the context they must be studied within. We propose two sets of experiments:

1. within context from 2*3 to 2*12 words before and after a given token, we use the form and the combination *form/RM*;
2. combinations of various features in a 2*7-word window size around a given token.

5.4. Evaluation

The evaluation is done through the comparison with the reference data. We compute precision, recall and F-measure at the level of categories (Sebastiani, 2002). The baseline corresponds to the use of the form and combination *form/RM* in the 2*7-word window. This is the basic information directly available in the text, within the average size context.

6. Results

6.1. Detection of enunciations with reformulations

For the detection of enunciations with reformulations, filters 1, 3 and 4 provide the best combination. Second rule, stating that when markers occur with disfluencies and primes there is no reformulation, is not efficient. Indeed, in spoken language, disfluencies can occur around reformulated segments. With filters 1, 3 and 4, the best precision obtained is 0.66%, which is higher than the inter-annotator agreement (0.617 in *ESLO1* and 0.526 in *ESLO2*).

6.2. Extraction of reformulated segments

Tables 2 and 3 present the results obtained during the step dedicated to the extraction of reformulated segments: we indicate precision, recall and F-measure for categories *O*, *SEG1* and *SEG2*.

Table 2 indicates the results when only forms and combination *form/RM* are used. The difference between the experiments is due to the variation of the context size from 2*3 to 2*12 tokens. This set of experiments gives the average value around 0.60: the average precision is around 0.70, while the average recall is around 0.58. Reformulated segments are detected with precision between 0.40 and 0.62, although the recall values are low (less than 0.20). The baseline is competitive. The best average values are observed with the context size with 2*11 and 2*4 words.

Table 3 indicates the results obtained with various combinations of features in the 2*7-word context. The experiments

form	*POS*	*chunkBI*	*chunk*	*heu*	*num*	*dmf*	*stem*	*RM*	*ref.*
...									
la	DET	I-PP	PP	N	28	MIL	la	O	O
cuisson	NC	I-PP	PP	N	29	MIL	cuisson	O	O
rapide	ADJ	I-PP	PP	N	30	MIL	rapid	O	O
quoi	PROWH	I-PP	PP	EUH	31	MIL	quoi	O	O
des	DET	B-NP	NP	N	32	MIL	de	O	SEG1
morceaux	NC	I-NP	NP	N	33	MIL	morceau	O	SEG1
nobles	ADJ	I-NP	NP	N	34	MIL	nobl	O	SEG1
ce	PRO	I-NP	NP	N	35	MIL	ce	O	O
qu’	PROREL	B-NP	NP	N	36	MIL	qu’	O	O
ils	CLS	B-NP	NP	N	37	MIL	il	O	O
appellent	V	B-VN	VN	N	38	MIL	appellent	O	O
quoi	PROWH	B-NP	NP	EUH	39	MIL	quoi	O	O
c’	CLS	B-NP	NP	N	40	MIL	c’	M	M
est	V	B-VN	VN	N	41	MIL	est	M	M
à	P	B-PP	PP	N	42	MIL	à	M	M
dire	VINF	I-PP	PP	N	43	MIL	dir	M	M
les	DET	B-NP	NP	N	44	MIL	le	O	SEG2
rosbifs	ADJ	I-NP	NP	N	45	MIL	rosbif	O	SEG2
les	DET	I-NP	NP	N	46	MIL	le	O	SEG2
biftecks	NC	I-NP	NP	N	47	MIL	bifteck	O	SEG2
et	CC	B-CONJ	CONJ	N	48	MIL	et	O	SEG2
tout	PRO	B-NP	NP	N	49	MIL	tout	O	SEG2
ça	PRO	B-NP	NP	N	50	MIL	ça	O	SEG2
...									

Table 1: Excerpt from the annotated data, with reformulation.

	O			*SEG1*			*SEG2*		
Context size	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
*E1 (2*3 words)*	0,81	0,97	0,88	0,61	0,13	0,21	0,50	0,13	0,20
*E2 (2*4 words)*	0,81	0,95	0,88	0,51	0,18	0,27	0,39	0,13	0,20
*E3 (2*5 words)*	0,81	0,97	0,88	0,57	0,13	0,21	0,45	0,12	0,19
*E4 (2*6 words)*	0,81	0,97	0,88	0,59	0,11	0,19	0,46	0,12	0,19
*E5 (2*7 words) - Baseline*	0,82	0,94	0,88	0,47	0,13	0,21	0,37	0,18	0,24
*E6 (2*8 words)*	0,81	0,98	0,88	0,62	0,12	0,20	0,51	0,11	0,18
*E7 (2*9 words)*	0,81	0,97	0,88	0,59	0,11	0,19	0,49	0,14	0,22
*E8 (2*10 words)*	0,81	0,97	0,88	0,61	0,13	0,21	0,48	0,15	0,22
*E9 (2*11 words)*	0,82	0,96	0,88	0,61	0,20	0,30	0,51	0,16	0,24
*E10 (2*12 words)*	0,81	0,97	0,88	0,61	0,13	0,21	0,46	0,15	0,22

Table 2: Use of the form and of the combination form/RM within 2*3 to 2*12 context size.

outperform the baseline. *E1* corresponds to the use of several features (*POS, chunk, heu, stem, RM*) within 2*7-word context, and of several combinations (*chunk/RM, POS/RM, stem/RM*) within 2*1-word context. Other experiments derive from *E1*. The experiments show that the best detection of reformulated segments (F-measure over 0.30) are *E4, E6* and *E8*. It seems that use of additional features, and removal of the *POS* and *heu* features are beneficial.

We can propose several additional observations:

- the *O* positions are well detected,
- detection of the source and target segments remains difficult and shows variable performance,
- among the best experiments, we count the baseline (use of forms within 2*7-word context and of combination *form/RM*) and the experiments based on various combinations of features (Table 3),
- 2*4, 2*7 and 2*11 words are among the optimal context sizes.

6.2.1. Discussion

We also processed separately the *ESLO1* and *ESLO2* corpora, and the non-consensual annotations (annotators $A1$ and $A2$). The models created on each dataset (*ESLO1*/$A1$, *ESLO1*/$A2$, *ESLO2*/$A1$, *ESLO2*/$A2$) have been applied on other datasets in order to study the portability of these models. These experiments indicate that:

- it is easier to detect reformulated segments in *ESLO1*,

	O			*SEG1*			*SEG2*		
Various combinations	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Baseline	0,82	0,94	0,88	0,47	0,13	0,21	0,37	0,18	0,24
E1	0.84	0.94	0.89	0.47	0.42	0.44	0.70	0.17	0.27
*E2 (E1 + 2*7 words(chunk/RM))*	0.83	0.36	0.50	0.12	0.59	0.20	0.27	0.52	0.36
*E3 (E2 + 2*7 words(POS/RM))*	0.85	0.19	0.32	0.36	0.43	0.39	0.15	0.93	0.26
*E4 (E3 + 2*7 words(stem/RM))*	0.87	0.36	0.51	0.20	0.65	0.31	0.21	0.69	0.32
*E5 (E4 + 2*7 words(début))*	0.83	0.46	0.60	0.27	0.50	0.35	0.16	0.55	0.25
*E6 (E5 - 2*7 words(POS))*	0.84	0.48	0.61	0.23	0.52	0.32	0.22	0.66	0.33
*E7 (E5 - 2*7 words(chunk))*	0.84	0.26	0.40	0.20	0.60	0.30	0.18	0.78	0.30
*E8 (E5 - 2*7 words(heu))*	0.85	0.39	0.53	0.24	0.45	0.32	0.20	0.81	0.32
*E9 (E5 - 2*7 words(stem))*	0.85	0.32	0.47	0.31	0.28	0.29	0.15	0.82	0.26
*E10 (E5 - 2*7 words(RM))*	0.83	0.42	0.56	0.17	0.47	0.25	0.22	0.69	0.34

Table 3: Various combinations of features within 2*7-word context. The experiment *E1* corresponds to the combination of several features (POS, chunk, heu, stem, RM) within 2*7-word context, and of chunk/RM, POS/RM, stem/RM within 2*1-word context. Other experiments derive from *E1*.

whatever the model (corpus or annotator). Indeed, *ESLO2* contains much longer enunciations, which makes the segment detection more difficult;

- models related to annotators also show various performance, although less important than those related to the influence of corpora;
- quite frequently, detection of the target segments *SEG2* is easier than detection of the source segments *SEG1*.

Experiments with consensual reference data and merged corpora slightly improve the results.

6.2.2. Analysis of errors

An analysis of the results indicates that often segments can be detected but with frontiers different from those expected by the reference data. In Figure 2, we present two examples: in A we can find the manual reference annotations and in B the automatically detected segments. Reformulated segments are in blue and underlined. We observe that in these examples, the automatically detected segments are larger than the reference annotations. Notice that several of the automatically proposed solutions are acceptable: similarly, during the manual annotation process, the annotators had several hesitations as for the right size of the segments. We assume that the automatically computed output can be used as basis for the manual annotation.
As indicated above, it can be easier to detect target segment that source segment (example (4)).

(4) *A. oui enfin par industriel je veux dire euh j'ai le côté commercial*
B. oui enfin par industriel je veux dire euh j'ai le côté commercial [ESLO1_ENT_002_C]
(*oh well by industrial I mean euh I have commercial dispositions*)

Another type of errors is observed when RMs are merged with one of the segments. Notice that this mainly happens with the marker *disons*, which is the less grammaticalized in the function of reformulation.

6.3. Significance

Going beyond the theoretical contribution to the studies on reformulations and paraphrases, there is also a practical significance of the proposed work. The methods proposed and the results provided by these methods can be useful for studying and comparing various corpora, both spoken and written, from the point of view of discursive and reformulation structures they use. For instance, in spoken language, it can be interesting to know in which discusison conditions speakers need to better explain thier ideas and to produce more reformulations. In addition, reformulations and paraphrases are related in the way that reformulations may provide paraphrases. As we already noticed in Section 1., paraphrases are very useful in several NLP applications such as information retrieval and extraction, textual entailment, machine translation, etc.

7. Conclusion

We have proposed a work on automatic filtering of enunciations with reformulations and on extraction of reformulated segments. The work shows two main original points:

- the reformulations are studied in syntagmatic structures, and more specifically, they can be found within the following structure:
  SEG1 RM SEG2
- the work is done with spoken corpora, in which reformulations are frequent and can be observed thanks to specific markers.

We proposed to use a rule-based system for filtering out enunciations without reformulations, and a CRF-based system for the automatic detection and extraction of reformulated segments. Various rules, features, their combinations and context sizes are tested.
When filtering out the enunciations without reformulations, we reach up to 0.66% precision, which is higher than the inter-annotator agreement. Among the best experiments proposed for the extraction of reformulated segments, we find the baseline (use of forms and of combination *form/RM*), and of various combinations of features

A. et cinq kilomètres c'est-à-dire j'avais quatre kilomètres à faire quatre et quatre huit je faisais huit kilomètres tous les jours et à pieds ah oui
B. et cinq kilomètres c'est-à-dire j'avais quatre kilomètres à faire quatre et quatre huit je faisais huit kilomètres tous les jours et à pieds ah oui [ESLO1_ENT_011_C]
(*and five kilometers in other words I had four kilometers to make four and four gives eight I was making height kilometers every day and by foot oh yes*)

A. et et vous par exemple approximativement vous combien de fois euh quelle est la fréquence avec laquelle vous regardez le dictionnaire c'est à dire une fois par mois une fois par an une fois par euh oh
B. et et vous par exemple approximativement vous combien de fois euh quelle est la fréquence avec laquelle vous regardez le dictionnaire c'est à dire une fois par mois une fois par an une fois par euh oh [ESLO1_ENT_047_C]
(*and and you for instance approximately how many times euh what is the frequency with which you use the dictionary in other words once a moth once a year once a euh oh*)

Figure 2: Analysis of the automatic detection of reformulated segments: difference in the detection of frontiers of the segments.

(*e.g. chunk/RM*, *pos/RM*, *stem/RM*) in 2*7-word context. The best average results reach up to 0.65 F-measure, 0.75 precision and 0.63 recall. We observe that the reformulated segments remain difficult to be detected correctly: F-measure is seldom higher than 0.30 for both segments.

8. Future Work

We have several directions for future work. We can test other classifiers, such as Long Short Term Memory (LSTM) (Hochreiterand and Schmidhuber, 1997), and other features and their combinations. Once stabilized, the models can be used for pre-annotating new data and preparing data for human annotation. We assume this may help the manual annotation. We plan also to use the models generated with these three markers on enunciations with other RMs (*e.g. ça veut dire, j'allais dire, notamment, autrement dit, en d'autres termes, en d'autres mots*) and other corpora. This will enable to study whether reformulations introduced by different markers have common regularities.
Other directions for future work consist of analyzing these data from other points of view. For instance, we can also study prosodic and acoustic features for improving the distinction between enunciations with and without reformulations (Section 5.2.). The hypothesis is that enunciations with reformulations may also show phonetic specificities. This step may involve a machine learning-based system as well, instead of the rule-based system. Study of reformulations in written corpora (discussion fora and journalistic news) is yet another perspective. Since the two spoken corpora exploited have been built with similar principles but with 40 year difference, this offers the possibility to perform a diachronic study of PMs. Finally, we plan to exploit the paraphrases generated in this work in other NLP applications, such as information retrieval and extraction.

9. Bibliographical References

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

Bannard, C. and Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *ACL*, pages 597–604.

Barzilay, R. and McKeown, L. (2001). Extracting paraphrases from a parallel corpus. In *ACL*, pages 50–57.

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):78–93.

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

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

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

Bouamor, H. (2012). *Étude de la paraphrase sous-phrastique en traitement automatique des langues*. Thèse de doctorat, Université Paris Sud, Paris.

Callison-Burch, C., Cohn, T., and Lapata, M. (2008). Parametric: An automatic evaluation metric for paraphrasing. In *COLING*, pages 97–104.

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

Dagan, I., Roth, D., Sammons, M., and Zanzotto, F. (2013). *Recognizing Textual Entailment*. Morgan & Claypool Publishers, Milton Keynes, UK.

Eshkol-Taravella, I. and Grabar, N. (2014). Paraphrastic reformulations in spoken corpora. In *POLTAL 2014, LNCS V8686*, pages 425–437.

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.

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

Fuchs, C. (1982). *La paraphrase*. PUF, Paris.

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

Gulich, E. and Kotschi, T. (1983). Les marqueurs de la reformulation paraphrastique. *Cahiers de linguistique française*, 5:305–351.

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

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

Hochreiterand, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 7(8):1735–1780.

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

Hwang, Y. (1993). "eh bien", "alors", "enfin" et "disons" en français parlé 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*, pages 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*, pages 145–153.

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

Lavergne, T., Cappé, O., and Yvon, F. (2010). Practical very large scale CRFs. In *ACL*, pages 504–513, July.

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

Madnani, N. and Dorr, B. J. (2010). Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36:341–387.

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

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

Milicevic, J. (2007). *La paraphrase : Modélisation de la paraphrase langagière*. Peter Lang.

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

Och, F. and Ney, H. (2000). Improved statistical alignment models. In *ACL*, pages 440–447.

Pasça, M. and Dienes, P. (2005). Aligning needles in a haystack: Paraphrase acquisition across the Web. In *IJCNLP*, pages 119–130.

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

Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Quirk, C., Brockett, C., and Dolan, W. (2004). Monolingual machine translation for paraphrase generation. In *EMNLP*, pages 142–149.

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

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.

Scarpa, F. (2010). *La Traduction spécialisée. Une approche professionnelle à l'enseignement de la traduction*. University of Ottawa Press. Language Arts & Disciplines, Ottawa, Canada.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

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

Shen, S., Radev, D., Patel, A., and Erkan, G. (2006). Adding syntax to dynamic programming for aligning comparable texts for the generation of paraphrases. In *ACL-COLING*, pages 747–754.

Shinyama, Y., Sekine, S., Sudo, K., and Grishman, R. (2002). Automatic paraphrase acquisition from news articles. In *Proceedings of HLT*, pages 313–318.

Tellier, I., Eshkol, I., Dupont, Y., and Wang, I. (2014). Peut-on bien chunker avec de mauvaises étiquettes POS? In *TALN 2014*.

Teston-Bonnard, S. (2008). "je veux dire" est-il toujours une marque de reformulation? In MC Le Bot, et al., editors, *Rivages linguistiques. La Reformulation. Marqueurs linguistiques. Stratégies énonciatives*, pages 51–69. PUR, Rennes.

Vila, M., Antònia Mart, M., and Rodríguez, H. (2011). Paraphrase concept and typology. a linguistically based and computationally oriented approach. *Procesamiento del Lenguaje Natural*, 46:83–90.

Vila, M., Rodréguez, H., and Martí, M. (2014). Relational paraphrase acquisition from wikipedia: The WRPA method and corpus. *Natural Language Engineering*, pages 1–35.