Extracting Lay Paraphrases of Specialized Expressions from Monolingual Comparable Medical Corpora

Louise Deléger
INSERM U872 Eq.20
Paris, F-75006 France
louise.deleger@spim.jussieu.fr

Pierre Zweigenbaum
CNRS, LIMSI
Orsay, F-91403 France
pz@limsi.fr

Abstract

Whereas multilingual comparable corpora have been used to identify translations of words or terms, monolingual corpora can help identify paraphrases. The present work addresses paraphrases found between two different discourse types: specialized and lay texts. We therefore built comparable corpora of specialized and lay texts in order to detect equivalent lay and specialized expressions. We identified two devices used in such paraphrases: nominalizations and neo-classical compounds. The results showed that the paraphrases had a good precision and that nominalizations were indeed relevant in the context of studying the differences between specialized and lay language. Neo-classical compounds were less conclusive. This study also demonstrates that simple paraphrase acquisition methods can also work on texts with a rather small degree of similarity, once similar text segments are detected.

1 Introduction

Comparable corpora refer to collections of texts sharing common characteristics. Very often comparable corpora consist of texts in two (or more) languages that address the same topic without being translations of each other. But this notion also applies to monolingual texts. In a monolingual context, comparable corpora can be texts from different sources (such as articles from various newspapers) or from different genres (such as specialized and lay texts) but dealing with the same general topic. Comparable corpora have been used to perform several Natural Language Processing tasks, such as extraction of word translations (Rapp, 1995; Chiao and Zweigenbaum, 2002) in a multilingual context or acquisition of paraphrases (Barzilay and Lee, 2003; Shinyama and Sekine, 2003) in a monolingual context. In this work[1], we are interested in using comparable corpora to extract paraphrases.

Paraphrases are useful to various applications, including information retrieval (Ibrahim et al., 2003), information extraction (Shinyama and Sekine, 2003), document summarization (Barzilay, 2003) and text simplification (Elhadad and Sutaria, 2007). Several methods have been designed to extract paraphrases, many of them dealing with comparable text corpora. A few paraphrase acquisition approaches used plain monolingual corpora to detect paraphrases, such as (Jacquemin, 1999) who detects term variants or (Pasca and Dienes, 2005) who extract paraphrases from random Web documents. This type of corpus does not insure the actual existence of paraphrases and a majority of methods have relied on corpora with a stronger similarity between the documents, thus likely to provide a greater amount of paraphrases. Some paraphrase approaches used monolingual parallel corpora, *i.e.* different translations or versions of the same texts. For instance (Barzilay and McKeown, 2001) detected paraphrases in a corpus of English translations of literary novels. However such corpora are not easily available and approaches which rely instead on other types of corpora are actively investigated.

Bilingual parallel corpora have been exploited for acquiring paraphrases in English (Bannard and Callison-Burch, 2005) and French (Max, 2008). Comparable corpora are another useful source of paraphrases. In this regard, only closely related corpora have been used, especially and almost exclusively corpora of news sources reporting the

[1] This paper is an extension of the work presented in (Deléger and Zweigenbaum, 2008a) and (Deléger and Zweigenbaum, 2008b), more specifically, a new corpus is added, an additional type of paraphrase (based on neo-classical compounds) is extracted and the evaluation is more relevant.

Proceedings of the 2nd Workshop on Building and Using Comparable Corpora, ACL-IJCNLP 2009, pages 2–10,
Suntec, Singapore, 6 August 2009. ©2009 ACL and AFNLP

same events. (Barzilay and Lee, 2003) generated paraphrase sentences from news articles using finite state automata. (Shinyama and Sekine, 2003) extracted paraphrases through the detection of named entities anchors in a corpus of Japanese news articles. In the medical domain, (Elhadad and Sutaria, 2007) worked with a comparable, almost parallel, corpus of medical scientific articles and their lay versions to extract paraphrases between specialized and lay languages.

We aim at detecting paraphrases in medical corpora in the same line as (Elhadad and Sutaria, 2007) but for French. This type of paraphrases would be a useful resource for text simplification or to help authoring medical documents dedicated to the general public. However, in a French medical context, it is difficult to obtain comparable corpora of documents with a high level of similarity, such as pairs of English scientific articles and their translations in lay language, or news articles reporting the same events used in general language (Barzilay and Lee, 2003; Shinyama and Sekine, 2003). Therefore, in addition to using this type of comparable corpora, we also tried to rely on corpora with less similarity but more easily available documents: lay and specialized documents from various sources dealing with the same overall medical topic.

We describe our experiment in building and exploiting these corpora to find paraphrases between specialized and lay language. Issues at stake involve: (i) how to collect corpora as relevant as possible (Section 2.1); (ii) how to identify passages which potentially convey comparable information (Section 2.2); and (iii) what sorts of paraphrases can be collected between these two types of discourse, which is addressed in Section 2.3, through the identification of two kinds of paraphrases: nominalization paraphrases and paraphrases of neo-classical compounds. An evaluation of the method (Section 2.4) is conducted and results are presented (Section 3) and discussed (Section 4).

2 Material and Methods

2.1 Building comparable corpora of lay and specialized texts

Today, a popular way of acquiring a corpus is collecting it from the Web (Kilgarriff and Grefenstette, 2003), as it provides easy access to an unlimited amount of documents. Here we focus on monolingual comparable corpora of specialized and lay medical French documents, with the objective of identifying correspondences between the two varieties of languages in these documents. We collected three corpora from the Web dealing with the following three topics: nicotine addiction, diabetes and cancer.

When dealing with a Web corpus several issues arise. The first one is the relevance of the documents retrieved to the domain targeted and is highly dependant on the method used to gather the documents. Possible methods include querying a general-purpose search engine (such as Google) with selected key words, querying a domain-specific search engine (in domains where they exist) indexing potentially more relevant and trustworthy documents, or directly downloading documents from known relevant websites. Another important issue specific to our type of corpus is the relevance to the genre targeted, *i.e.* lay vs. specialized. Hence the need to classify each collected document as belonging to one genre or the other. This can be done by automatic categorisation of texts or by direct knowledge of the sources of documents. In order to obtain a corpus as relevant as possible to the domain and to the genres, we used direct knowledge and restricted search for selecting the documents. In the case of the cancer topic, we had knowledge of a website containing comparable lay and specialized documents: the Standards, Options: Recommandations website[2] which gives access to guidelines on cancer for the medical specialists on the one hand and guides for the general public on the same topics on the other hand. This case was immediate: we only had to download the documents from the website. This corpus is therefore constituted of quite similar documents (professional guidelines and their lay versions). The other two corpora (on nicotine addiction and diabetes), however, were built from heterogeneous sources through a restricted search and are less similar. We first queried two health search engines (the health Web portals CISMeF[3] and HON[4]) with key words. Both allow the user to search for documents targeted to a population (*e.g.*, patient-oriented documents). We also queried known relevant websites for documents dealing with our chosen topics. Those were

[2] http://www.sor-cancer.fr/

[3] http://www.cismef.org/

[4] http://www.hon.ch/

French governmental websites, including that of the HAS[5] which issues guidelines for health professionals, and that of the INPES[6] which provides educational material for the general public; as well as health websites dedicated to the general public, including Doctissimo[7], Tabac Info Service[8], Stoptabac[9] and Diabète Québec[10].

The corpus dealing with the topic of diabetes served as our development corpus for the first type of paraphrases we extracted, the other two corpora were used as test corpora.

Once collected, a corpus needs to be cleaned and converted into an appropriate format to allow further processing, *i.e.* extracting the textual content of the documents. HTML documents typically contain irrelevant information such as navigation bars, footers and advertisements—referred to as "boilerplate"—which can generate noise. Boilerplate removal methods can rely on HTML structure, visual features (placement and size of blocks) and plain text features. We used HTML structure (such as meta-information and density of HTML tags) and plain text (such as spotting phone and fax numbers and e-mails, as often appear at the end of documents) to get rid of boilerplate.

2.2 Aligning similar text segments

We hypothesize that paraphrases will be found more reliably in text passages taken from both sides of our comparable corpora which address similar topics. So, as a first step, we tried to relate such passages. We proceeded in three steps:

1. as multiple topics are usually addressed in a single text, we performed topic segmentation on each text using the TextTiling (Hearst, 1997) segmentation tool. A segment may consist of one or several paragraphs;

2. we then tried to identify pairs of text segments addressing similar topics and likely to contain paraphrases. For this we used a common, vector-based measure of text similarity: the cosine similarity measure which we computed for each pair of topic segments in the cross-product of both corpus sides (each segment was represented as a bag of words);

3. we selected the best text segment pairs, that is the pairs with a similarity score equal or superior to 0.33, a threshold we determined based on the results of a preliminary study (Deléger and Zweigenbaum, 2008a).

2.3 Extracting paraphrases

We are looking for paraphrases between two varieties of language (specialized and lay), as opposed to any kind of possible paraphrases. We therefore endeavoured to determine what kind of paraphrases may be relevant in this regard. A common hypothesis (Fang, 2005) is that specialized language uses more nominal constructions where lay language uses more verbs instead. We test this hypothesis and build on it to detect specialized-lay paraphrases around noun-to-verb mappings (a first version of this work was published in (Deléger and Zweigenbaum, 2008b)). A second hypothesis is that medical language contains a fair proportion of words from Latin and Greek origins, which are referred to as neo-classical compounds. The meaning of these words may be quite obscure to non-experts readers. So one would expect to find less of these words in lay texts and instead some sort of paraphrases in common language. We therefore tried to detect these paraphrases as a second type of specialized vs. lay correspondences.

2.3.1 Paraphrases of nominalizations

A first type of paraphrases we tried to extract was paraphrases between nominal constructions in the specialized side (such as *treatment of the disease*) and verbal constructions in the lay side (such as *the disease is treated*). This type of paraphrases involves nominalizations of verbal phrases and is built around the relation between a deverbal noun (*e.g. treatment*) and its base verb (*e.g. treat*). Therefore, we relied on a lexicon of French deverbal nouns paired with corresponding verbs (Hathout et al., 2002) to detect such pairs in the corpus segments. These noun-verb pairs served as anchors for the detection of paraphrases. In order to design paraphrasing patterns we extracted all pairs of deverbal noun and verb with their contexts from the development corpus. The study of such pairs with their contexts allowed us to establish a set of lexico-syntactic paraphrasing patterns[11]. An example of such patterns can be seen in Table 1.

[5] http://www.has-sante.fr/

[6] http://www.inpes.sante.fr/

[7] http://www.doctissimo.fr/

[8] http://www.tabac-info-service.fr/

[9] http://www.stop-tabac.ch/

[10] http://www.diabete.qc.ca/

[11] Texts were first tagged with Treetagger (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/).

Specialized	**Lay**
N_1 PREP (DET) N_2	V_1 (DET) N_2
N_1 PREP (DET) $N_2 A_3$	V_1(DET) $N_2 A_3$
N_1 A_2	V_1(DET) N_2

Table 1: Example paraphrasing patterns (a shared index indicates equality or synonymy. N=noun, V=verb, A=adjective, PREP=preposition, DET=determiner, 1 in index = pair of deverbal noun and verb)

The general method was to look for corresponding content words (mainly noun and adjective) in the contexts. We defined corresponding words as either equal or synonymous (we used lexicons of synonyms as resources[12]). Equals may have either the same part-of-speech, or different parts-of-speech, in which case stemming[13] is performed to take care of derivational variation (*e.g.*, *medicine* and *medical*). We then applied the patterns to both development and test corpora.

The patterns thus designed are close to the transformation rules of (Jacquemin, 1999) who detects morpho-syntactico-semantic variants of terms in plain monolingual corpora. One difference is that our patterns are built around one specific type of morphological variation (noun to verb variation) that seemed relevant in the context of the specialized/lay opposition, as opposed to any possible variation. We also identify the paraphrases by comparing the two sides of a comparable corpus while (Jacquemin, 1999) starts from a given list of terms and searches for their variants in a plain monolingual corpus. Finally, we do not apply our method on terms specifically but on any expression corresponding to the patterns.

2.3.2 Paraphrases of neo-classical compounds

We then extracted paraphrases of neo-classical compounds as a second type of paraphrases that seemed relevant to the opposition between lay and specialized languages. This means that we looked for neo-classical compounds on one side of the corpora and equivalents in modern language on the other side. To do this we relied on the morphosemantic parser DériF (Namer and Zweigenbaum, 2004). DériF analyzes morphologically complex words and outputs a decomposition of those words into their components and a definition-like gloss of the words according to the meaning of the components in modern language when they are from Greek or Latin origins. For instance the French word *gastrite* (*gastritis*) is decomposed into *gastr+ite* and its gloss is *inflammation de l'estomac* (*inflammation of stomach*).

We first ran the analyzer on the specialized side of the corpora to detect neo-classical compounds. Then we searched for paraphrases of those compounds based on the output of DériF, that is we looked for the modern-language equivalents of the word components (in the case of *gastritis* this means searching for *inflammation* and *stomach*) close to each other within a syntactic phrase (we empirically set a threshold of 4 words as the maximum distance between the modern-language translations of the components). A pattern used to search those paraphrases is for instance:

C → ((DET)? N PREP)? (DET)? C_1 W$^{0-4}$ C_2

where C is a neo-classical compounds in a specialized text segment, C_1 and C_2 are the modern-language components of C, N is a noun, PREP a preposition, DET a determiner and W an arbitrary word.

2.4 Evaluation

We first evaluated the quality of the extracted paraphrases by measuring their precision, that is, the percentage of correct results over the entire results. We computed precision for each type of paraphrases.

We then estimated recall for the first type of paraphrases (nominalization paraphrases): the percentage of correct extracted paraphrases over the total number of paraphrases that should have been extracted. We used as gold standard a random sample of 10 segment pairs from which we manually extracted paraphrases.

Finally, since we aim at detecting paraphrases between lay and specialized languages, we also looked at the relevance of the two types we chose to extract. That is, we evaluated the coherence of the results with our two initial hypotheses, which are expected to apply when both a specialized text segment and a lay text segment convey similar information: (1) nominalizations are more often used in specialized texts while lay texts tend to

[12] The lexicons used came from the Masson and Robert dictionaries.

[13] Stemming was performed using the Lingua::Stem perl package (http://search.cpan.org/~snowhare/Lingua-Stem-0.83) which is similar to the Snowball stemmers (http://snowball.tartarus.org)

		Specialized		Lay
(a)	N_s	...the benefits of *smoking cessation*...	N_l	...withdrawal symptoms of *smoking cessation*...
(b)	N_s	...regular *use of tobacco* concerned...	N_l	...*tobacco use* is the first cause...
(c)	N_s	...which goes with *smoking cessation*...	V_l	...who wants *to stop smoking*...

Table 2: Sample cases used to compute the conditional probability for nominalizations; (a) and (b) represent cases where a paraphrase was expected but did not occur and (c) a case where a paraphrase was indeed used. N = nominalization; V = verbal form.

		Specialized		Lay
(a)	C_s	...*glycemia* is lower...	C_l	...a drop of *glycemia*...
(b)	C_s	...the starting point of *thrombosis*...	C_l	...the risk of *thrombosis*...
(c)	C_s	...especially *cardiopathies* and...	M_l	...25% of *heart diseases*...

Table 3: Sample cases used to compute the conditional probability for neo-classical compounds; (a) and (b) represent cases where a paraphrase was expected but did not occur and (c) a case where a paraphrase was indeed used. C = compound; M = modern.

replace them with verbs; (2) specialized texts use more neoclassical compounds while lay texts give a paraphrase in modern language.

To evaluate (1) we measured the conditional probability $P(V_l|N_s)$ that a nominalization pattern N_s in a specialized segment be replaced by a matching verbal pattern V_l in a corresponding lay segment. These patterns are the paraphrasing patterns defined in Section 2.3.1 and exemplified in Table 1. Table 2 gives examples of cases taken into account when computing this probability, *i.e.* cases where both text segments convey the same information, as a nominalization in the specialized side and as a nominalization or a verbal paraphrase in the lay side. Formally, the probability can be estimated by $\frac{|Par_{N_s \to V_l}|}{|ExpPar_{N_s \to V_l}|}$, where $|Par_{N_s \to V_l}|$ is the number of correct extracted paraphrases involving a nominalization in a specialized segment and a verbal construction in the corresponding lay segment (case (c) of Table 2), and $|ExpPar_{N_s \to V_l}|$ the expected number of paraphrases. The expected number of paraphrases corresponds to the total number of instances where a specialized text segment contains a nominalization and the corresponding lay segment conveys the same information, expressed either as a nominalization or as a paraphrasing verbal construction (cases (a), (b) and (c) of Table 2). It is therefore computed as the sum of $|Par_{N_s \to V_l}|$ and $|Par_{N_s \to N_l}|$, the latter referring to the number of occurrences where both the specialized and lay segments match the same nominalization pattern, *i.e.*, instances where a paraphrase was expected but did not occur (cases (a) and (b) of Table 2). For instance *use of tobacco* on one side and *tobacco use* on the other side, as in (b), is a case where one would have expected a paraphrase such as *tobacco is used*. Note that matching allows the same flexibility as described in Section 2.3.1 in terms of synonyms and morphological variants. To test whether this tendency of using verbal constructions instead of nominalizations is indeed stronger in lay texts we also measured the reverse, *i.e.* the conditional probability $P(V_s|N_l)$, given a nominalization pattern N_l in a lay segment, that it be replaced with a matching verbal pattern V_s in the corresponding specialized segment, computed as $\frac{|Par_{N_l \to V_s}|}{|ExpPar_{N_l \to V_s}|}$. If our hypothesis is verified, this reverse probability should be lower then the direct probability.

In the same way, to evaluate (2) we measured the conditional probability $P(M_l|C_s)$ that a neo-classical compound C_s in a specialized segment be replaced by a modern-language equivalent M_l in a corresponding lay segment. Table 3 gives examples of cases taken into account when computing this probability, that is cases where both text segments convey the same information, as a neo-classical compound in the specialized side and as a neo-classical compound or a modern-language paraphrase in the lay side. Formally, it can be estimated by $\frac{|Par_{C_s \to M_l}|}{|ExpPar_{C_s \to M_l}|}$, where $|Par_{C_s \to M_l}|$ is the number of correct extracted paraphrases involving a neo-classical compound in a specialized

	Diabetes		Nicotine addiction		Cancer	
	S	L	S	L	S	L
docs	135	600	62	620	22	16
words	580,712	461,066	595,733	603,257	641,584	228,742
segment pairs	183		547		438	

Table 4: Sizes of the corpora (Number of documents, words and segment pairs; S=specialized, L=lay)

	Diabetes	**Nicotine add.**	**Cancer**
total paraph.	42	79	93
correct paraph.	30	62	62
precision	71.4%	78.5%	75.8%

Table 5: Precision for nominalization paraphrases (at the type level, not token level)

	Diabetes	**Nicotine add.**	**Cancer**
total paraph.	39	3	3
correct paraph.	24	3	3
precision	61.5%	100%	100%

Table 6: Precision for paraphrases of neo-classical compounds (at the type level, not token level)

segment and a modern-language equivalent in the corresponding lay segment (case (c) of Table 3) , and $|ExpPar_{C_s \to M_l}|$ is the expected number of paraphrases (case (a), (b) and (c) of Table 3). The expected number of paraphrases is the sum of $|Par_{C_s \to M_l}|$ and $|Par_{C_s \to C_l}|$, the latter referring to the number of occurrences where both the specialized and lay segments contains the same neo-classical compound (instances where a paraphrase was expected but did not occur, for instance cases (a) and (b) of Table 3). We then measured the reverse, *i.e.* the conditional probability $P(M_s|C_l)$, given a neo-classical compound C_l in a lay segment, that it be replaced with a modern-language equivalent M_s in the corresponding specialized segment, computed as $\frac{|Par_{C_l \to M_s}|}{|ExpPar_{C_l \to M_s}|}$.

3 Results

Table 4 gives size figures for each side (lay and specialized) of the three corpora in terms of documents, words and segment pairs.

Evaluation of the quality of the extracted paraphrases shows that precision is rather good for both type of paraphrases (see Tables 5 and 6), although the figures cannot be considered signicative for paraphrases of compounds extracted in the tobacco and cancer corpora given the small number of paraphrases (only 3 paraphrases in both cases).

Examples of nominalization paraphrases and paraphrases of neo-classical compounds are given in Tables 7 and 8. The last line of Table 7 shows an example of incorrect paraphrase, which is due to the synonymy link established between French words *charge* and *poids* which is not valid in that particular context. The last line of Table 8 also gives an incorrect example, which is caused by the imprecision of the modern-language paraphrase which is only partially equivalent to the neo-classical compound.

Specialized	**Lay**
consommation régulière *regular use*	consommer de façon régulière *to use in a regular fashion*
gêne à la lecture *reading difficulty*	empêche de lire *prevents from reading*
évolution de l'affection *evolution of the condition*	la maladie évolue *the disease is evolving*
*prise en charge *the taking care of*	prendre du poids *to take on weight*

Table 7: Examples of extracted nominalization paraphrases (* indicates an incorrect example)

With regard to the quantitative evaluation of the nominalization paraphrases, we measured a 30% recall on our sample of segment pairs, meaning that out of the 10 manually extracted paraphrases only 3 were automatically detected by our method. Cases of non-detected paraphrases were due to the restrained scope of the paraphrasing patterns, as well as to the presence of synonyms not contained

Specialized	Lay
leucospermie *leucospermia*	Augmentation du nombre de globules blancs dans le sperme *Increase in the number of white cells in the sperm*
glycémie *glycemia*	la quantité de sucre dans le sang *amount of sugar in the blood*
prostatectomie *prostatectomy*	l'ablation de la prostate *ablation of the prostate*
*hyperglycémie *hyperglycemia*	le taux de sucre dans le sang *proportion of sugar in the blood*

Table 8: Examples of extracted paraphrases of neo-classical compounds (* indicates an incorrect example)

in our lists.

Table 9 displays results for the investigation on the coherence of our first initial hypothesis that specialized texts use nominalizations where lay texts use verbal constructions. The conditional probability that a nominalization be replaced with a verbal construction is higher for nominalizations in specialized texts than for the reverse direction, which means that nominalizations in specialized texts are indeed more likely to be replaced by verbal constructions in lay texts than nominalizations in lay texts by verbal constructions in specialized texts. Results for the second hypothesis (neo-classical compounds in specialized texts tend to be replaced by modern-language equivalents in lay texts) are given in Table 10. As for the first hypothesis, the conditional probability for the neo-classical compounds in the specialized texts is higher, which seems to be coherent with the initial hypothesis. However, given the very small number of paraphrases, we cannot draw a significative conclusion as regards this second type of paraphrases.

4 Discussion

In this work we built comparable corpora of specialized and lay texts on which we implemented simple paraphrase acquisition methods to extract certain types of paraphrases that seemed relevant in the context of specialized and lay language: paraphrases based on nominalization vs. verbal constructions and paraphrases based on neo-classical compounds vs. modern-language expressions. The precision measured on the set of detected paraphrases is rather good, which indicates good quality of the paraphrases (hence of the paraphrasing patterns and extracted segments).

An originality of this work lies in the fact that, in contrast to approaches working with more closely related comparable corpora (Barzilay and Lee, 2003; Shinyama and Sekine, 2003; Elhadad and Sutaria, 2007), we also gathered comparable corpora of documents which, although addressing the same general topics (nicotine addiction, diabetes), were a priori rather different since coming from various sources and targeted to different populations. We showed that simple paraphrase acquisition methods could also work on documents with a lesser degree of similarity, once similar segments were detected. Indeed the precision of the extracted paraphrases is within the same range for the three corpora we built, despite the fact that one corpus (the cancer corpus) was composed of more similar documents than the other two.

We extracted a type of paraphrases much less exploited in existing work, with the exception of (Elhadad and Sutaria, 2007), that is paraphrases between specialized and lay language. This meant that we had to take into account what kind of paraphrases might be relevant, therefore the methods used to extract them were more constrained and supervised than approaches aiming at detecting any type of paraphrases. We based a part of our work on the hypothesis that among relevant types were paraphrases involving nominalizations of verbal contructions, meaning that lay texts tend to use verb phrases where specialized texts use deverbal noun contructions. Our results seem to support this hypothesis. Such paraphrases therefore seem to be interesting advice to give to authors of lay texts. Future work includes testing our method on English and comparing the results for the two languages. We would expect them to be fairly similar since the tendency to use nominal constructions in scientific literature has also been observed for English (Fang, 2005). The second part of our work exploited the hypothesis that lay texts use modern-language expressions where specialized texts use neo-classical compound words. In this case, the paraphrases were too few to enable us to draw a significative conclusion. Testing this method on different and larger corpora might give more insight into the relevance of extracting this type of paraphrases. As it is, this work is still experimental and needs to be further investigated.

	Diabetes		**Nicotine addiction**		**Cancer**	
	S→L	L→S	S→L	L→S	S→L	L→S
# paraphrases ($\|Par_{N_s \to V_l}\|$ or $\|Par_{N_l \to V_s}\|$)	44	37	140	76	73	57
# expected paraphrases ($\|ExpPar_{N_s \to V_l}\|$ or $\|ExpPar_{N_l \to V_s}\|$)	712	695	1675	1626	770	772
Conditional Probability ($P(V_l\|N_s)$ or $P(V_s\|N_l)$)	0.062	0.053	0.084	0.047	0.095	0.074

Table 9: Conditional probability for nominalization paraphrases in both directions, specialized-lay (S→L) and lay-specialized (L→S)

	Diabetes		**Nicotine addiction**		**Cancer**	
	S→L	L→S	S→L	L→S	S→L	L→S
# paraphrases ($\|Par_{C_s \to M_l}\|$ or $\|Par_{C_l \to M_s}\|$)	53	40	18	0	3	0
# expected paraphrases ($\|ExpPar_{C_s \to M_l}\|$ or $\|ExpPar_{C_l \to M_s}\|$)	686	675	196	178	1482	1479
Conditional Probability ($P(M_l\|C_s)$ or $P(M_s\|C_l)$)	0.074	0.059	0.092	0	0.002	0

Table 10: Conditional probability for paraphrases of neo-classical compounds in both directions

Its major drawback is the low number of paraphrases, in particular for the paraphrases of neo-classical compounds which brought inconclusive results. In order to gain insight on the low quantity of paraphrases of neo-classical compounds, we manually looked at sample text segments from the nicotine addiction and cancer corpora (the two corpora where very few paraphrases were extracted) and could not find any paraphrase of neo-classical compounds. This would seem to indicate that the low quantity of this type of paraphrases is due to the characteristics of the corpora rather than to defects of our extraction technique. As for the nominalization paraphrase, even though the method brought more paraphrases and gave encouraging results, their quantity is still quite small. The recall computed on a sample of segment pairs is low. This is mainly due to the fact that we set up rather rectricted paraphrasing patterns. This was done to ensure a high precision but caused the recall to fall. A future step would be to improve recall by modifying some aspects of the paraphrasing patterns while trying to keep a good precision.

Regardless of recall, the number of nominalization paraphrases in itself is also small. This can be due to the fact that we restrict ourselves to one specific type of paraphrases, but also to the facts that we first align and select similar text segments, that the coverage of our corpora might not be sufficient, and that we work on comparable corpora of lesser similarity than other methods. Future work to increase the number of paraphrases involves using clusters of text segments instead of pairs, increasing the corpus sizes and developing methods to detect other types of paraphrases besides the two kinds investigated here.

5 Conclusion

We presented a method based on comparable medical corpora to extract paraphrases between specialized and lay languages. We identified two kinds of paraphrases, nominalization paraphrases and paraphrases of neo-classical compounds, the first type seeming to indeed reflect some of the systematic differences between specialized and lay texts while the second type brought too few results to draw a signicative conclusion.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *HLT-NAACL*, pages 16–23, Edmonton, Canada.

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *ACL/EACL*, pages 50–57.

Regina Barzilay. 2003. *Information Fusion for Multidocument Summarization: Paraphrasing and Generation*. Ph.D. thesis, Columbia University.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for French-English translations in comparable medical corpora. In *Proc AMIA Symp*, pages 150–4.

Louise Deléger and Pierre Zweigenbaum. 2008a. Aligning lay and specialized passages in comparable medical corpora. In *Stud Health Technol Inform*, volume 136, pages 89–94.

Louise Deléger and Pierre Zweigenbaum. 2008b. Paraphrase acquisition from comparable medical corpora of specialized and lay texts. In *Proceedings of the AMIA Annual Fall Symposium*, pages 146–150, Washington, DC.

Noemie Elhadad and Komal Sutaria. 2007. Mining a lexicon of technical terms and lay equivalents. In *ACL BioNLP Workshop*, pages 49–56, Prague, Czech Republic.

Zhihui Fang. 2005. Scientific literacy: A systemic functional linguistics perspective. *Science Education*, 89(2):335–347.

Nabil Hathout, Fiammetta Namer, and Georgette Dal. 2002. An Experimental Constructional Database: The MorTAL Project. In *Many Morphologies*, pages 178–209.

Marti A. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Ali Ibrahim, Boris Katz, and Jimmy Lin. 2003. Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings of the second international workshop on Paraphrasing*, pages 57–64, Sapporo, Japan. Association for Computational Linguistics.

Christian Jacquemin. 1999. Syntagmatic and paradigmatic representations of term variation. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 341–348, College Park, Maryland.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–47.

Aurélien Max. 2008. Local rephrasing suggestions for supporting the work of writers. In *Proceedings of GoTAL*, Gothenburg, Sweden.

Fiammetta Namer and Pierre Zweigenbaum. 2004. Acquiring meaning for French medical terminology: contribution of morphosemantics. In Marius Fieschi, Enrico Coiera, and Yu-Chuan Jack Li, editors, *MEDINFO*, pages 535–539, San Francisco.

Marius Pasca and Peter Dienes. 2005. Aligning needles in a haystack: Paraphrase acquisition across the web. In *Proceedings of IJCNLP*, pages 119–130.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 320–322.

Yusuke Shinyama and Satoshi Sekine. 2003. Paraphrase acquisition for information extraction. In *Proceedings of the second international workshop on Paraphrasing (IWP)*, pages 65–71, Sapporo, Japan.