Identifying Medical Paraphrases in Scientific versus Popularization Texts in French for Laypeople Understanding

Ioana Buhnila[1]
[1]LiLPa UR 1339 - Linguistique, Langues, Parole, University of Strasbourg, France
ioana.buhnila@etu.unistra.fr

Abstract

Scientific medical terms are difficult to understand for laypeople due to their technical formulas and etymology. Understanding medical concepts is important for laypeople as personal and public health is a lifelong concern. In this study, we present our methodology for building a French lexical resource annotated with paraphrases for the simplification of monolexical and multiword medical terms. In order to find medical paraphrases, we automatically searched for medical terms and specific lexical markers that help to paraphrase them. We annotated the medical terms, the paraphrase markers, and the paraphrase. We analysed the lexical relations and semantico-pragmatic functions that exists between the term and its paraphrase. We computed statistics for the medical paraphrase corpus, and we evaluated the readability of the medical paraphrases for a non-specialist coder. Our results show that medical paraphrases from popularization texts are easier to understand (62.66%) than paraphrases extracted from scientific texts (50%).

1 Introduction

Understanding medical terms is a challenge for laypeople of all ages and education level. In this study, we concentrated on adults that are not professionals of the medical field but are interested in understanding medical knowledge and research. Medical language is difficult to understand due to, partially, the large number of medical terms. The *term* represents a lexical unit that expresses concepts specific to a field of knowledge, recognised and shared by members of a community of specialists (Costa, 2005). The *term* belongs to an autonomous "subsystem" of the language with the goal of communicating technical or scientific knowledge (Contente, 2005). Medical terms are particularly difficult to understand because of their Greek and/or Latin etymology (Grabar and Hamon, 2015). They can be composed of a mix of prefixes/suffixes from these two ancient languages together with morphemes of the modern language. Laypeople have difficulties in understanding the meaning of medical terms such as "cholecystectomy", which is formed with two Greek basis, "chole" (=bile) and "ectomy" (=surgical removal), and in the middle of these, a Latin basis, "cystis" (=bladder) (Grabar and Hamon, 2015). We can simplify medical terms by using synonyms from the common language, but it is sometimes difficult to find the right synonym. In this paper, we explore the medical paraphrases as means of simplification of medical terms in French. *Paraphrasing* is the process of rewriting in order to explain or simplify a word, sentence, or phrase, while keeping the same meaning.

In this paper, we worked on scientific medical texts in French that treat a certain medical concept (diseases, treatments, medical procedures) and on their versions written for laypeople. We looked for paraphrases of these concepts created with simpler words and expressions from the common language. We evaluated the level of difficulty of these medical paraphrases for adult lay readers.

The annotated paraphrases will constitute a corpus of medical paraphrases that could be used as a textual resource in Natural Language Processing (NLP) and deep learning tasks.

Section 2 presents the medical corpus exploited and in Section 3 we describe the methodology we used to identify medical terms and paraphrase markers, and thus, the medical paraphrase. We continue with the annotation process in Section 4. Section 5 describes the evaluation of the medical paraphrases and their readability level according to a non-expert coder whose native language is

Proceedings of the Third Workshop on Scholarly Document Processing, pages 69–79
October 12–17, 2022.

French. We conclude with the potential use of our annotated corpus for scientific medical term simplification (in Section 6).

2 Related Work

In this section, we present several studies on the themes that our research is related to: paraphrases, medical paraphrase corpus, paraphrase markers, medical terms, and automatic paraphrase identification in French.

2.1 Paraphrases

In linguistics, *paraphrasing* represents the process of rewriting in order to explain or simplify a concept or phrase. There are multiple studies on the concept of paraphrase (Gühlich and Kotschi, 1983; Fuchs, 1994; Rossari, 1990; Vassiliadou, 2013; Grabar and Eshkol-Taravella, 2016; Eshkol-Taravella and Grabar, 2017; Steuckardt, 2018; Fuchs, 2020; Pennec, 2020; Vassiliadou, 2020), from which we highlight:

- The concept of *paraphrasing* as the process of preserving the meaning and intending to get close to a semantic equivalence (Fuchs, 2020; Pennec, 2020; Vassiliadou, 2020);
- *Subphrastic paraphrase* (Bouamor, 2012), composed of words or groups of words that are semantically tied and are integrated in a sentence;
- *Subphrastic paraphrasing,* defined as the process of intra-lingual translation (translation with elements of the same language system, keeping the same meaning) that does not exceed the length of a sentence (our definition);
- *The classical paraphrase*, which expresses an equivalence based on a common semantic core (Fuchs, 1982; Bouamor, 2012; Kampeera, 2013; Pennec, 2020).

In this study, we chose to work on the large concept of *paraphrasing*, as our goal was to identify the largest sequence of words that are semantically equivalent. As we searched for paraphrases that coexist with the medical term in the same sentence, we worked exclusively on *subphrastic paraphrases*. We looked for any paraphrase that can be used to explain and simplify medical terms. The goal of our project was to build a corpus of medical paraphrases that can be used as a database for simplifying scientific medical concepts and adapting medical knowledge to laypeople (Cardon, 2021; Grabar and Hamon, 2015; Grabar and Hamon, 2016).

2.2 Paraphrase Markers

The classical way of identifying paraphrases is through specific markers. *The paraphrase markers* are linguistic elements that help to identify paraphrases in texts. They can be lexical, grammatical, or orthographic markers or cues of paraphrase (Fuchs, 2020; Steuckardt, 2018). Several studies on French focused on paraphrase markers based on the verb *dire* (to say), such as *c'est-à-dire* (that is), *ça veut dire* (that means), *pour dire autrement* (to say otherwise), *autrement dit* (otherwise said) (Vassiliadou, 2013; Grabar and Eshkol-Taravella, 2016; Steuckardt, 2018; Magri, 2018). These markers can have a narrative or paraphrastic role. Vassiliadou (2013) considers the marker *c'est-à-dire* (that is) as the typical paraphrase marker. Grabar and Eshkol-Taravella (2016) worked on specific markers for lexical paraphrases (*c'est-à-dire* (that is), *disons* (let's say), *ça veut dire* (it means)) using a rule-based system and manual annotations. Their study aimed at automatically classifying phrases with and without paraphrases. To identify paraphrases, Grabar and Eshkol-Taravella (2016) looked for the syntagmatic structure "S1 marker S2", where S1 is the paraphrased element and S2 is the paraphrase. These two parts are linked by the paraphrase markers cited above. Their study was conducted on two general oral language corpora and a medical forum corpus.

In our work, we also took into consideration the possibility of the *absence* of the paraphrase marker. We looked for *paraphrase cues* specific to the medical domain, as a scientific and specialized type of text. We classified the *paraphrase cues* into three types:

- *General language cues* that, through their semantics and use in discourse, refer to the simplification, definition, or explanation of concepts: *définition* (definition), *défini/e* (defined), etc.;
- *Grammatical cues* that announce a list of hyponyms of the medical term: *comme* (such as), *par exemple* (for example);

- *Cues specific to the medical domain* which are hypernyms of the medical terms: *maladie* (disease), *affection* (affection), *trouble* (disorder).

We manually analysed the corpus to find more paraphrases without markers or lexical cues. We found other markers, such as the *typographical cues* (parentheses or commas) (Steuckardt, 2018). Unlike Grabar and Eshkol-Taravella (2016), who worked on medical forum texts (which contain text that are very similar to oral written speech), we analysed written medical texts (scientific and popular articles) in order to create a set of sentences that contain medical paraphrases in natural language context (and not only in a lexicon). For this purpose, we used markers analysed in other similar works, but we also added additional markers and cues of paraphrase, presented in section 3.2.

2.3 Scientific Medical Terms and their Paraphrases

In order to locate the paraphrase, we first identified the medical term that is paraphrased. The aim of paraphrasing medical terms is to propose a meaning equivalent to the sequence of words from the common language, adapted to non-specialist readers, such as patients, students, or laypeople in general (Leroy et al., 2013; Brouwers et al., 2012; Pecout, Tran and Grabar, 2019).

Several different methods were experimented to identify medical terms and their paraphrases, for example searching for Latin or Greek prefixes and suffixes, using medical ontologies (Grabar and Hamon, 2016) or with term detection tools with n-gram patterns (Buhnila, 2018). Grabar and Hamon (2016) searched for medical terms in a corpus of Wikipedia articles using medical terminologies (Snomed International (Côté, 1996) and the French part of UMLS (*Unified Medical Language System*) (Donald et al., 1993). Their study focused on paraphrases that appear in free contexts, meaning that the technical terms and their paraphrases can be separated by several words. In the same study, they used the French morphological analyser DériF (Namer, 2009) to extract words in modern French from medical terms of Greek or Latin origin. For example, the term “myocardique” contains the modern French words “muscle” / muscle (myo) and “cœur” / heart (carde). The authors looked for these words in the corpora and extracted 2,596 definitory contexts automatically.

In this paper, we focused on simple and multiword medical terms and we used the SNOMED-3.5VF medical ontology (Côté, 1996) for scientific term extraction.

2.4 Medical Paraphrase Corpus for French

We can mention the study of Cardon and Grabar (2021) on 4,596 pairs of parallel sentences extracted from the CLEAR corpus (Grabar and Cardon, 2018), a medical corpus of popularization and scientific texts. The goal of the study was to automatically simplify biomedical texts using neural networks. Cardon and Grabar (2021) used several resources: the parallel phrases of the CLEAR corpus, a lexicon that matches complex medical terms with paraphrases easy to understand to laypeople (7,580 paraphrases for 4,516 medical terms) and 297,494 parallel sentences in the common language from WikiLarge (Zhang and Lapata, 2017). The WikiLarge corpus was automatically translated from English to French. Their experiments proved that using a medical lexicon of paraphrases and medical simplified phrases helped simplify biomedical texts.

The goal of our study was to build an annotated corpus of sentences that contain medical paraphrases in a natural language context and that can be used for the simplification of medical texts and scientific medical concepts. We present our method in Section 3.

3 Methodology

For this study, we worked on the CLEAR corpus which is composed of French scientific medical texts and medical texts adapted for laypeople (Grabar and Cardon, 2018). Our method consisted of automatically identifying simple and multiword medical terms with the SIFR-BioPortal annotator (Tchechmedjiev et al., 2018). We also tested other annotators for French, such as Bio-YODIE (Gorrell et al., 2018) and PyMedTermino (Lamy et al., 2015), but SIFR-BioPortal proved to be the most intuitive to use. SIFR-BioPortal works by parsing texts for medical terms from the SNOMED-3.5VF medical ontology (Côté, 1996) (released by ASIP Santé). This ontology contains 150,906 scientific medical concepts in French. In order to identify specific markers for the medical domain, we looked for words that collocate with a term and the relations that this term may have with other elements of the sentence. More precisely, we run a

Perl script to identify relation markers (Condamines, 2018) that link a medical term to its paraphrase (Ramadier, 2016), such as hypernymy, hyponymy, synonymy, meronymy.

We expected to find paraphrases in the context of the medical term in the same sentence. After the automatic identification of the medical terms and paraphrase markers, we manually annotated the sentences to find out whether the paraphrases are correct or not. We also annotated the paraphrases for the lexical relations and semantico-pragmatic functions, such as definition, explanations, etc. (presented in detail in section 4.2 and 4.3). We present each step of the methodology in detail as it follows.

3.1 Corpus of Study

Our corpus of study was the CLEAR Cochrane corpus, which is a part of the CLEAR corpus (Grabar and Cardon, 2018). CLEAR is a comparable corpus composed of *scientific texts from the medical field* designed for experts and *simplified texts* written for laypeople. The texts were written by researchers of the Cochrane Foundation. Grabar and Cardon (2018) collected a number of 8,789 texts in November 2017, of which 3,815 were duplicates of the same medical concept: asthma, arthritis, motor neuron disease, etc. The expert corpus contains 2,840,003 tokens and the laypeople corpus counts 1,515,051 tokens.

CLEAR Cochrane	N° of texts	Same theme texts	Size (token)
Expert (EX)	8,789	3,815	2,840,003
Laypeople (GP)			1,515,051

Table 1: Size of the CLEAR Cochrane corpus by text type (Grabar and Cardon, 2018).

The CLEAR Cochrane corpus is built with comparable texts on the same theme, where a scientific text is followed by its simplified version. For our study, we decided to separate expert and laypeople texts in two sub-corpora: scientific corpus written for experts (CLEAR EX) and general public corpus (CLEAR GP). Our hypothesis is that scientific texts have more medical terms while general public texts contain more synonyms, paraphrases, or explanations in the common language. We split the texts into sentences using end-of-line characters (. ; ! ; ?) to display one sentence per line. Once the corpus was cleaned and aligned, we proceeded to automatically identify the medical terms (see Table 1).

3.2 Automatic Annotation of Medical Terms and Paraphrase Markers

We identified the medical terms in our corpus with the help of a Perl script and the French version of the SIFR-BioPortal annotator (Tchechmedjiev et al., 2018). The annotator provides 28 medical terminologies in French. We chose the SNOMED-3.5VF ontology because it contains a wide variety of medical concepts: administrative and treatments, agents, anatomy, diagnoses, drugs, symptoms, disease, procedures, etc. This large panel of medical concepts and the search by lemma helped us tag a large number of medical terms in our corpus of study.

As for the *paraphrase markers*, we listed the most frequent ones from the literature, to which we added markers according to our own observations from the corpus:

- Markers formed on the French verb *dire* (to say) (*c'est-à-dire* (it means), *ça veut dire* / *veut dire* (meaning), *pour dire autrement* (to say otherwise), *autrement dit* (in other words) (Vassiliadou, 2013; Vassiliadou, 2016; Grabar and Eshkol-Taravella, 2016; Steuckardt, 2018; Magri, 2018);

- Markers derived from the verbs *désigner* (to designate) and *signifier* (to signify) (Péry-Woodley and Rebeyrolle, 1998; Charolles and Coltier, 1986);

- Markers derived from the verb *être* (to be) with its different morphological forms, *est un/une/des* (is a), *sont un/une/des* (are a/some) (Meyer, 2001; Grabar and Hamon, 2016) followed by hypernyms from the medical domain such as "disease", "affection" and "disorder";

- Markers that are specific to our corpora, such as the ones formed on the verb *appeler* (to call) (*qu'on appelle, ce que l'on appelle* (what it's called), *est aussi appelé / aussi appelé* (is also called / also called) and others, such as *doit être compris comme* (must be understood as), *au sens de* (in the sense of).

These paraphrase markers are domain-independent (except medical hypernyms) and can indicate different types of relations between the medical term and its paraphrase (further details in Section 4.2).

4 Paraphrase Annotation Process

In this section we present different levels of the annotation process of the medical paraphrases. This annotation was manually done in order to assess the quality of the paraphrases that were automatically identified with previous tasks. In this paper we annotated the status of the paraphrase, the lexical relations and the semantico-pragmatic relations that exists between the medical term and its paraphrase.

4.1 Status of the Paraphrase

We chose five different possible values for the status of the paraphrase, as follows:

- *yes*: the sentence contains a correct paraphrase;
- *yes<rev>*: the sentence contains a reversed paraphrase (the paraphrase is found before the medical term);
- *yes<2+>*: there are two or more correct paraphrases in the same sentence;
- *yes<2+><rev>*: there are two or more correct paraphrases in the same sentence, with at least one reversed paraphrase;
- *no*: the sentence does not contain a correct paraphrase.

4.2 Lexical Relations

We classified the lexical relations that exist between the paraphrase and the corresponding medical term: synonymy, hyponymy, hypernymy, meronymy. Medical hypernyms (Săpoiu, 2013) have an important role in the classification of scientific medical concepts (e.g. “scrub typhus”) into wide classes that are easier to understand for laypeople, such as “bacterial disease” (Grabar and Hamon, 2015). For instance, in the case of hyponymy, the term “antibiotics” is the hypernym, and the paraphrase simplifies the meaning of the term by using hyponyms such as “chloramphenicol, tetracycline and doxycycline”.

4.3 Semantico-pragmatic Functions

The semantico-pragmatic functions express the reasons that motivate the writer to use paraphrases (such as definition, designation, exemplification, explanation, rephrasing) (Eshkol-Taravella and Grabar, 2017). In this study, we adapted this taxonomy, originally created on oral texts of common language, to written texts in the medical domain. We defined these functions as follows:

- *Definition*: the term is given a definition because it is considered to be too technical or domain specialised, thus difficult to understand;
- *Designation*: the term is paraphrased using another word or term;
- *Exemplification*: the paraphrase is a list of examples (several entities of the same type) that help to illustrate the meaning of the term;
- *Explanation*: the term is explained through a particular situation or procedure;
- *Rephrasing*: the meaning of the term is expressed with simpler words;

Definitive contexts are marked by specific lexical cues: *définition* (definition), *défini/e* (defined), *défini/e comme* (defined as). The phrases *tel/lle/s/lles que* (such as) and *par exemple* (for example) announce the paraphrase through an exemplification.

4.4 Readability Level of Paraphrases

Paraphrases can be easier or more difficult to understand by laypeople. The complexity is given by the use of technical words. For instance, the medical term “antibiotics” could be simpler to understand than “chloramphenicol”. In this sense, we asked a coder who is not a specialist in the field of medicine to evaluate a sample of correct medical paraphrases. We evaluated the level of comprehension of paraphrases through the manual annotation of a sample of correct paraphrases. We selected a sample of 300 paraphrases that were labelled as correct by our two coders (both not specialists in medicine), 150 from each type of corpus (scientific and for laypeople). We evaluate the comprehension of the paraphrases by three levels:

- *Level 1* – easy to understand: the paraphrase is easier to understand than the term (there are words from the common language in the paraphrase);

- *Level 2* – same complexity: same level of complexity or technicity between the term and the paraphrase, meaning that both the term and the paraphrase are difficult to understand;

- *Level 3* – difficult to understand: the paraphrase is more complex or technical than the term.

The annotation is done by a French native speaker, who is studying Linguistics at a Masters 2 degree level. The student annotated the paraphrases identified by the other coder of the study (ourselves). We present the results of this evaluation in the next section.

5 Results and Data Evaluation

The automatic extraction of the sentences containing both the medical terms annotated by SIFR-BioPortal and occurrences of markers or paraphrase indicators is done with Perl scripts. We adapted our scripts to identify all morphological forms and to automatically annotate medical terms and markers/cues. We obtained 4,681 sentences for the corpus of scientific texts (CLEAR EX) and 3,975 sentences for the corpus of medical texts for the general public (CLEAR GP). These sentences were therefore analysed manually by two coders. We present the results and statistics of these annotations in the tables below.

5.1 Coder Agreement

We computed the agreement between two coders, ourselves, and a French native speaker, Master-level student. We computed the Kappa annotator agreement (Cohen, 1960), the precision and recall of paraphrases identified. We show in Table 2 and 3 the number of paraphrases that were identified as correct medical paraphrases by both coders, the number of paraphrases that received the same "status" tag ("yes", "yes-rev", "no"), in both corpora. We also computed the number of paraphrases tagged differently by both coders. We decided to not include "yes<2+>" and "yes<2+><rev>" tags in this study, as these paraphrases appear in a small number. We will analyse them in future studies.

CLEAR EX		
Statistics	**Coder 1**	**Coder 2**
Paraphrases with ***yes***	1321	1714
Paraphrases with ***yes-rev***	37	50
Paraphrases with ***no***	3323	2917
Different tag paraphrases - ***total***	**948**	
Same tag paraphrases - ***yes***	1059	
Same tag paraphrases - ***yes-rev***	7	
Same tag paraphrases - ***no***	2667	
Same tag paraphrases - ***total***	**3733**	
Total number of paraphrases	**4681**	

Table 2: Coder data statistics on CLEAR EX

As for the general public corpus, we analysed only the annotated sentences (1,903 out of 3,975). We calculated the precision, the recall, and the relative frequencies in order to interpret data equally.

CLEAR GP		
Statistics	**Coder 1**	**Coder 2**
Paraphrases with ***yes***	671	707
Paraphrases with ***yes-rev***	55	22
Paraphrases with ***no***	1177	1174
Different tag paraphrases - ***total***	**291**	
Same tag paraphrases - ***yes***	552	
Same tag paraphrases - ***yes-rev***	17	
Same tag paraphrases - ***no***	1043	
Same tag paraphrases - ***total***	**1612**	
Total number of paraphrases	**1903**	

Table 3: Coder data statistics on CLEAR GP

We calculated the recall as the number of paraphrases tagged with "yes" or "no" by both coders, divided by the number of paraphrases tagged with "yes" or "no" by coder 1 and respectively by coder 2. We considered one annotation as the gold standard and then we changed the other way around (in Tables 4 and 5, the recall is computed with coder 1, and coder 2 respectively, as reference).

$$Recall = \frac{common\ paraphrases\ Coder1\&\ Coder2}{paraphrases\ Coder1}$$

Figure 1: Coder recall formula

CLEAR EX		
Measures	**Coder 1**	**Coder 2**
Precision - *yes*	0.29	0.38
Precision - *yes - average*	**0.34**	
Precision - *no*	0.71	0.62
Precision - *no - average*	**0.67**	
Precision - *same tag*	**0.80**	
Recall - *yes*	0.78	0.60
Recall - *yes - average*	**0.69**	
Recall - *no*	0.80	0.91
Recall - *no - average*	**0.86**	
Recall - *total average*	**0.78**	
Kappa annotator score	**0.55**	

Table 5: Statistics on CLEAR EX

CLEAR GP		
Measures	**Coder 1**	**Coder 2**
Precision - *yes*	0.38	0.38
Precision - *yes - total*	**0.38**	
Precision - *no*	0.62	0.62
Precision - *no - average*	**0.62**	
Precision - *same tag*	**0.85**	
Recall - *yes*	0.84	0.80
Recall - yes - *total*	**0.82**	
Recall - *no*	0.89	0.89
Recall - *no - average*	**0.89**	
Recall - *total average*	**0.86**	
Kappa annotator score	**0.68**	

Table 7: Statistics on CLEAR GP

The big differences in the number of "yes" tag paraphrases were due to different decisions of the coders, as the coder 1 decided not to consider *abbreviations* as *paraphrases,* while the coder 2 considered them as paraphrases. We intend to automatically annotate abbreviations in future studies for further analysis and conduct new analysis with and without abbreviations as paraphrases. Results proved that precision, recall, and *Cohen's Kappa* annotator are higher for the general public corpus than for the expert corpus. We also used the *ReCal tool* (Freelon, 2013) to do ordinal, interval, and ratio-level scores on both annotations. We gave numeric values to our tags, 1 for "yes", 2 for "yes-rev" and 3 for "no". The highest agreement score was the ordinal one, with **0.707** for the general public corpus and of **0.566** for the expert corpus.

We assume that these score differences were due to the higher level of technicity of the expert corpus, thus making it more difficult to assess the same tags for the paraphrases by both coders, while in the general public corpus the paraphrases were easier to analyse and evaluate.

Data	**CLEAR EX**	**CLEAR GP**
File size	23405 bytes	9515 bytes
N° coders	2	2
N° cases	4681	1903
N° decisions	9362	3806

Table 4: Corpus data for the ReCal Tool

ReCal Tool	**EX**	**GP**
Measures	**Score**	**Score**
Krippendorff's alpha (nominal)	0.552	0.688
Krippendorff's alpha (ordinal)	**0.566**	**0.707**
Krippendorff's alpha (interval)	0.565	0.705
Krippendorff's alpha (ratio)	0.562	0.701

Table 6: Measure scores obtained with ReCal

We analysed the absolute and relative frequencies of lexical relations and semantico-pragmatic functions for both corpora. We compared the average relative frequencies of both annotations, and we observed that the lexical relation of hypernymy is the most frequent in both corpora with a score of **63.32%** for the expert corpus and a score of **62.39%** for the general public corpus. We observed that the semantic-pragmatic function of definition had similar scores (**49.95%** and **52.28%** respectively). This can be justified by the fact that the definitory context has, most of the time, the following syntax:

medical term – paraphrase marker – medical hypernym – paraphrase

Semantico-pragmatic functions	**CLEAR EX**			**CLEAR GP**		
	A.F C1	**A.F C2**	**Av R.F**	**A.F C1**	**A.F C2**	**Av R.F**
Definition	723	342	**49.95 %**	356	239	**52.28 %**
Designation	30	152	8.53 %	18	83	8.87 %
Exemplifica-tion	242	222	**21.76 %**	128	113	**21.17 %**
Explanation	28	209	11.11 %	30	97	11.15 %
Rephrasing	43	141	8.63 %	37	37	6.50 %
N° phrases	**1066**		**100%**	**569**		**100%**

Table 8: Lexical relations between medical terms and their paraphrases (A.F=absolute frequency; Av R.F=average relative frequency; C1=coder 1; C2=coder 2; N° phrases: phrases with “yes” and “yes-rev” in common for both coders)

In the example: *La bronchectasie est une maladie respiratoire chronique* (Bronchiectasis is a chronic respiratory disease), the term “La bronchectasie” is paraphrased in a definitory sentence introduced by the medical hypernym “une maladie”.

Lexical relations	**CLEAR EX**			**CLEAR GP**		
	A.F C1	**A.F C2**	**Av R.F**	**A.F C1**	**A.F C2**	**Av R.F**
Synonymy	86	162	11.63 %	57	83	12.30 %
Hyponymy	245	218	**21.71 %**	128	114	**21.26 %**
Hypernymy	668	682	**63.32 %**	339	371	**62.39 %**
Meronymy	67	4	3.33%	45	1	4.04 %
N° phrases	**1066**		**100%**	**569**		**100%**

Table 9: Semantico-pragmatic functions between medical terms and their paraphrases (A.F=absolute frequency; Av R.F=average relative frequency; C1=coder 1; C2=coder 2; N° phrases: phrases with “yes” and “yes-rev” in common for both coders)

We observed the same situation for the lexical relation of hyponymy and the semantico-pragmatic function of exemplification, as they have almost the same scores (21.71% and 21.76% for CLEAR EX and 21.26% and 21.17% for CLEAR GP), meaning that they were annotated as appearing in the same context.

5.2 Complexity for Laypeople

The manual annotation of the level of comprehension of paraphrases showed that paraphrases from CLEAR GP are easier to understand (**62.66%** in comparison with **50%** for the scientific corpus). Meanwhile, the number of opaque paraphrases (where the paraphrase is as difficult to understand as the medical term because few words from the common language are used) is higher in the scientific corpus (**42%** compared to **27.33%** for the simplified version). This can be explained by the bigger number of scientific terms used as paraphrases in the expert texts.

Level	**CLEAR EX**		**CLEAR GP**	
	Abs F	**Rel F**	**Abs F**	**Rel F**
1: Easy to understand	75	**50%**	94	**62.66 %**
2: Same level of complexity	63	42%	41	27.33 %
3: Difficult to understand	12	8%	15	10%
Paraphrases	**150**		**150**	

Table 10: Assessment of the level of comprehension of medical paraphrases (Abs F=absolute frequency; Rel F=relative frequency)

6 Conclusion and further research

Our study has shown that medical paraphrases are present in both scientific and popularization texts. There is a higher number of paraphrases in the general corpus and are also easier to understand, both for annotation tasks and for lay readers comprehension. The analysis and evaluation of lexical relations and semantico-pragmatic functions that can be identified between the medical term and its paraphrase highlighted relations such as hyponymy and hyponymy help to identify more correct paraphrases. The same result is observed with the semantico-pragmatic functions of definition and exemplification. In further studies we will also conduct quantitative and qualitative analyses of paraphrase markers (or their absence) and compare them in scientific and popular texts. We could also evaluate the level of readability of each type of lexical relation and semantico-pragmatic function and assess which type of simplifications are easier to understand for laypeople. Further analysis could focus on whether the identified paraphrases are scientifically accurate and allow laypeople to be correctly informed about medical topics.

Here we created a corpus of **1,635 paraphrases of scientific medical terms in French** and an **annotated corpus of 6,584 phrases** that contain scientific medical terms and paraphrase markers. Once the annotation process is finished, the annotated corpora will be shared with the scientific community on the github repository. We are currently using the corpus for Natural Language Processing (NLP) tasks such as generating medical paraphrases for scientific terms and binary classification with deep learning and neural networks such as *OpenNMT* (Klein et al., 2020) and *APT* (Adversarial Paraphrasing Task) (Nighojkar et Licato, 2021).

Our method and experiences can also be applied on other Romance languages close to French, such as Romanian (Buhnila, 2021). Our corpus of medical paraphrases can constitute a useful lexical resource for scientific medical texts simplification system for adult lay readers or patients.

References

Houda Bouamor. 2012. *Etude de la paraphrase sous-phrastique en traitement automatique des langues*. Université Paris Sud - Paris XI. Français. ⟨NNT: 2012PA112100⟩. ⟨tel-00717702⟩.

Laetitia Brouwers, Delphine Bernhard, Anne-Laure Ligozat and Thomas François. 2012. *Simplification syntaxique de phrases pour le français (syntactic simplification for french sentences) [in French]. In Actes de la conférence conjointe JEP-TALN-RECITAL 2012, 2 : TALN* :211–224. Grenoble, France : ATALA/AFCP.

Ioana Buhnila. 2018. *Simplification lexicale entre les textes scientifiques et les textes de vulgarisation du domaine de la médecine.* Mémoire de Master, Université de Strasbourg, Strasbourg, France.

Ioana Buhnila. 2021. *Building a Corpus of Medical Paraphrases in Romanian. In Proceedings of the The 16th Edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing – ConsILR-2021*, Iasi, 139-152.

Rémi Cardon. 2021. *Simplification automatique de textes techniques et spécialisés. Informatique et langage [cs.CL].* Université de Lille. Français. ⟨NNT: 2021LILUH007⟩. ⟨tel-03343769v2⟩.

Rémi Cardon and Natalia Grabar. 2021. *Simplification automatique de textes biomédicaux en français : lorsque des données précises de petite taille aident (French Biomedical Text Simplification : When Small and Precise Helps). In Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, 275–277, Lille, France. ATALA.

Michel Charolles and Danielle Coltier. 1986. Le contrôle de la compréhension dans une activité rédactionnelle : l’exemple des paraphrases paraphrastiques. Pratiques 49 (1): 51 66. https://doi.org/10.3406/prati.1986.2450.

Jacob Cohen. 1960. *A coefficient of agreement for nominal scales*. Educ. Psychol. Meas., 20, 27-46.

Condamines, A. 2018). Nouvelles perspectives pour la terminologie textuelle. J. Altmanova; M. Centrella; K.E. Russo. Terminology and Discourse, Peter Lang, 1-13. 978-3-0343-2415-1. ff10.3726/978-3-0343-2414-4ff. ffhalshs-01899150f.

Madalena Contente. 2005. *Termes et textes : la construction du sens dans la terminologie médicale. Actes des septièmes Journées scientifiques du réseau de chercheurs Lexicologie Terminologie Traduction*, 453 65. Bruxelles, Belgique.

Rute Costa. 2005. *Texte, terme et contexte. Actes des septièmes Journées scientifiques du réseau de chercheurs Lexicologie Terminologie Traduction*, 79-88. Bruxelles, Belgique.

Réné Côté. 1996. *Répertoire d’anatomopathologie de la SNOMED internationale*, v3.4. Université de Sherbrooke, Sherbrooke, Québec.

Deen Freelon. 2013. *ReCal OIR: Ordinal, Interval, and Ratio Intercoder Reliability as a Web Service. International Journal of Internet Science. 8 (1)*, 10-16.

Iris Eshkol-Taravella and Natalia Grabar. 2017. *Taxinomie dans les paraphrases du point de vue de la linguistique de corpus.* Syntaxe et Sémantique, vol. 18, no. 1, 149-184.

Iris Eshkol-Taravella and Natalia Grabar. 2018. *Paraphrases : de l’étude outillée dans les corpus disponibles vers leur détection automatique*. Langages N° 212 (4), 5-16.

Catherine Fuchs. 1982. *La Paraphrase*. PUF, Paris, 184 pages.

Catherine Fuchs. 1994. *Paraphrase et énonciation*. Editions OPHRYS, 185 pages.

Catherine Fuchs. 2020. Paraphrase et paraphrase : un chassé-croisé entre deux notions. Autour de la paraphrase, 36, Droz, Coll. Recherches et Rencontres, 978-2-600-06051-6, 41-55.

Genevieve Gorrell, Xingyi Song and Angus Roberts. 2018. Bio-YODIE: A Named Entity Linking System for Biomedical Text. arXiv:1811.04860 [cs], http://arxiv.org/abs/1811.04860 , p. 1-5.

Natalia Grabar and Iris Eshkol-Taravella. 2016. *Disambiguation of occurrences of paraphrase markers c’est-à-dire, disons, ça veut dire. JADT 2016 : 13ème Journées internationales d’Analyse statistique des Données Textuelles*, 1-13. Nice, France.

Natalia Grabar and Thierry Hamon. 2015. *Extraction automatique de paraphrases grand public pour les termes médicaux. 22ème Traitement Automatique des Langues Naturelles*, 14. Caen, France.

Natalia Grabar and Thierry Hamon. 2016. *Exploitation de la morphologie pour l'extraction automatique de paraphrases grand public des termes médicaux. Traitement Automatique des Langues, Varia*, 57 (1), 85 109.

Natalia Grabar and Rémi Cardon. 2018. CLEAR - Simple Corpus for Medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA). Tilburg, the Netherlands: Association for Computational Linguistics*. https://doi.org/10.18653/v1/W18-7002, 3–9.

Elisabeth Gühlich and Thomas Kotschi. 1983. *Les marqueurs de la paraphrase paraphrastique*. Cahiers de Linguistique française 5, 305-351.

Wannachai Kampeera. 2013. *Analyse linguistique et formalisation pour le traitement automatique de la paraphrase*. Linguistique. Université de Franche-Comté. Français. ⟨NNT: 2013BESA1011⟩. ⟨tel-01288926⟩.

Guillaume Klein, François Hernandez, Vincent Nguyen, and Jean Senellart. 2020. *The OpenNMT Neural Machine Translation Toolkit*: 2020 Edition. AMTA 2020.

Anaïs Koptient and Natalia Grabar. 2020. *Rated Lexicon for the Simplification of Medical Texts. The Fifth International Conference on Informatics and Assistive Technologies for Health-Care*, Medical Support and Wellbeing HEALTHINFO 2020, Porto, Portugal.

Jean-Baptiste Lamy, Alain Venot and Catherine Duclos. 2015. PyMedTermino: an open-source generic API for advanced terminology services. *Studies in Health Technology and Informatics*, IOS Press, 2015, 210, pp.924-8. ⟨hal-03650024⟩.

Gondy Leroy, David Kauchak and Mouradi Obay. 2013. *A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty*. Int J Med Inform, 82(8), 717–730.

Donald A. B. Lindberg, Betsy Humphreys and Alexa Mccray. 1993. *The Unified Medical Language System*, Methods Inf Med, vol. 32, no 4, 281-291.

Mounica Maddela, Fernando Alva-Manchego and Wei Xu. 2020. *Controllable text simplification with explicit paraphrasing*. arXiv preprint arXiv:2010.11004.

Véronique Magri. 2018. *Marqueurs de paraphrase : exploration outillée et contrastive dans deux corpus narratifs*. Langages N° 212 (4), 35-50.

Ingrid Meyer. 2001. *Extracting Knowledge-Rich Contexts for Terminography: A conceptual and methodological framework*. Dans D. Bourigault, C. Jacquemin, & M.-C. L'Homme (Éds), Recent Advances in Computational Terminology, 279-302, Amsterdam: John Benjamins.

Fiammetta Namer. 2009. *Morphologie, Lexique et TAL : l'analyseur DériF*. TIC et Sciences cognitives, Hermes Sciences Publishing, London.

Animesh Nighojkar and John Licato. 2021. *Improving Paraphrase Detection with the Adversarial Paraphrasing Task*. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7106–7116, Online. Association for Computational Linguistics.

Anaïs Pecout, Tran, Thi M. and Natalia Grabar. 2019. *Améliorer la diffusion de l'information sur la maladie d'Alzheimer : étude pilote sur la simplification de textes médicaux.* Ela. Etudes de linguistique appliquée N° 195 (3): 325 41.

Blandine Pennec. 2020. *Les paraphrases : des formes méta-énonciatives par excellence. Spécificités et introducteurs*. Autour de la paraphrase, 36, Droz, Coll. Recherches et Rencontres, 57-75.

Marie-Paule Péry-Woodley and Josette Rebeyrolle. 1998. *Domain and genre in sublanguage text: definitional microtexts in three corpora*, LREC, 987-992.

Lionel Ramadier. 2016. *Indexation et apprentissage de termes et de relations à partir de comptes rendus de radiologie*. Informatique. Université Montpellier, Français. ⟨NNT: 2016MONTT298⟩. ⟨tel-01479769v2⟩.

Corinne Rossari. 1990. *Projet pour une typologie des opérations de paraphrase*. Cahiers de linguistique française 11, 345-359.

Camelia Săpoiu. 2013. *Hiponimia în terminologia medicală.* Modalităţi de abordare în semantică şi lexicografie. Piteşti, Editura Trend, 199 pages.

Agnès Steuckardt. 2018. *Les marqueurs de paraphrase formés sur dire : exploration outillée*. Langages N° 212 (4), 17-34.

Andon Tchechmedjiev, Amine Abdaoui, Vincent Emonet, Stella Zevio et Clement Jonquet. 2018. *SIFR annotator: ontology-based semantic annotation of French biomedical text and clinical notes.* BMC bioinformatics, 19(1), 405.

Hélène Vassiliadou. 2013. C'est-à-dire (que) : embrayeur d'énonciation. Semen. Revue de sémio-linguistique des textes et discours, no 36 (octobre). 1-14. http://journals.openedition.org/semen/9684. https://doi.org/10.4000/semen.9684.

Hélène Vassiliadou. 2016. *Mouvements de réflexion sur le dire et le dit : c'est-à-dire, autrement dit, ça* *veut dire.* Histoires de dire. Petit glossaire des marqueurs formés sur le verbe dire, L. Rouanne & J.-C. Anscombre (éds), Bern/Berlin/Bruxelles/New York/Oxford/Wien, Peter Lang, 339-364.

Hélène Vassiliadou. 2020. *Peut-on aborder la notion de "paraphrase" autrement que par la typologie des marqueurs ?* Pour une analyse sémasiologique et onomasiologique. Olga Inkova. Autour de la Paraphrase, Droz, 978-2-600-06051-6, 77-94.

Seid M. Yimam and Chris Biemann. 2018. *Par4Sim-Adaptive Paraphrasing for Text Simplification*. arXiv preprint arXiv:1806.08309.

Xingxing Zhang and Mirella Lapata. 2017. *Sentence simplification with deep reinforcement learning*. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 584–594, Copenhagen, Denmark: Association for Computational Linguistics.