# LARGEMED: a Resource for Identifying and Generating Paraphrases for French Medical Terms

Ioana Buhnila, Amalia Todirascu

# LARGEMED: a Resource for Identifying and Generating Paraphrases for French Medical Terms

**Ioana Buhnila, Amalia Todirascu**
ATILF UMR 7118 (CNRS-University of Lorraine), LiLPa UR 1339 (University of Strasbourg)
Nancy, Strasbourg (France)
ioana.buhnila@univ-lorraine.fr, todiras@unistra.fr

**Abstract**

This article presents a method extending an existing French corpus of paraphrases of medical terms RefoMed (Buhnila, 2023) with new data from Web archives created during the Covid-19 pandemic. Our method semi-automatically detects new terms and paraphrase markers introducing paraphrases from these Web archives, followed by a manual annotation step to identify paraphrases and their lexical and semantic properties. The extended large corpus LARGEMED could be used for automatic medical text simplification for patients and their families. To automatise data collection, we propose two experiments. The first experiment uses the new LARGEMED dataset to train a binary classifier aiming to detect new sentences containing possible paraphrases. The second experiment aims to use correct paraphrases to train a model for paraphrase generation, by adapting T5 Language Model to the paraphrase generation task using an adversarial algorithm.

**Keywords:** medical terms, paraphrases, automatic paraphrase generation

## 1. Introduction

Text adaptation aims to produce a simplified version (for example at lexical level) of the original document for a specific target audience with reading difficulties or insufficient knowledge. In the medical domain, text adaptation for patients or patients' families helps them to better understand their illness and to better fight against it. Medical knowledge is shared by health specialists and experts, but lay people have difficulties to understand the content of medical texts, due to the high density of scientific terms with opaque meaning. Terms are lexical units identifying a concept from a specialised domain (Condamines, 1997). Thus, text adaptation systems propose synonyms, alternative explanations, definitions or paraphrases of difficult medical terms for the target audience (patients or people with shallow medical knowledge).

However, automatic text adaptation requires large corpora or paraphrase datasets. Few French NLP resources are available for the medical domain, such as the parallel medical corpus **CLEAR**, containing aligned scientific and simplified medical abstracts (Grabar and Cardon, 2018), or the **RefoMed** dataset (Buhnila, 2023) containing pairs of medical terms and their paraphrases.

Thus, we propose a method for building a large corpus, containing medical terms and their various paraphrases, useful for automatic text simplification. **Paraphrases** are considered to be sequences of words aiming to preserve the sense of the paraphrased term (Fuchs, 1982; Vassiliadou, 2020), with various surface forms: simple words, phrases, sentences. Building such datasets is a difficult task, due to the various lexical and syntactic forms of the paraphrases. In this article, we adopt the definition proposed by Eshkol-Taravella and Grabar (2017): we consider that definitions, exemplifications and explanations represent various forms of **subsentential paraphrases** (paraphrases identified in the same sentence as the term). We aim to build a large resource with various forms of subsentential paraphrases for medical terms that might enhance the accessibility of medical knowledge to a non-specialist audience.

In this context, we propose two main contributions: **(1)** a large corpus **LARGEMED**, containing French terms and their subsentential paraphrases semi-automatically extracted from medical texts. The resource is annotated with lexical relations and semantico-pragmatic functions of the paraphrases; **(2)** some experiments aiming to extend LARGEMED by automatic paraphrase classification and generation;

Firstly, we present the concept of paraphrase in linguistics and NLP followed by our own definition. We continue with the state-of-the-art methods of classification and paraphrase generation, as well as the few French NLP resources available for medical domain used for automatic paraphrase classification, generation or text adaptation. Then, we describe the data found in the RefoMed corpus and the annotation guidelines applied to our own corpus containing Covid-19 terms and their paraphrases. In the next section we detail our method to collect data from Web archives used to complete RefoMed. Subsequently, we detail the classification and the generation experiments, based on LARGEMED, in order to eventually collect more data. We discuss our results and



*Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 141–151

conclude with future perspectives for our work.

## 2. Background

No unique definition of the notion of paraphrase is available in linguistics, computational linguistics and NLP. Fuchs (1982; 2020) considers that the paraphrase should be semantically equivalent to the paraphrased word or term. Eshkol-Taravella and Grabar (2017) adopt a broader point of view of the concept of paraphrase, assuming that it can have various lexical or syntactic forms while preserving similar or same meaning. Between the terms and their paraphrases, several lexical relations could be established: *synonymy*, *hypernymy*, *hyponymy* (2). Eshkol-Taravella and Grabar (2017) assume that the intention behind the usage of paraphrases in discours can exhibit several semantico-pragmatic functions, such as *definition* (1), *explanation*, *exemplification*, or *rephrasing*. We illustrate this linguistic variety with some examples extracted from the CLEAR corpus (Grabar and Cardon, 2018) (where the medical term is in **bold** font and the paraphrase in *italic*):

1. **Les troubles de l'équilibre** étaient définis si *le patient n'était pas en mesure de rester au moins cinq secondes en appui unipodal*.

   (**The equilibrium troubles** are defined as *the patient is not able to stay at least 5 seconds in single-leg support*.)

2. Les autres **traitements immunosuppresseurs** (*mycophénolate mofétil, cyclophosphamide, méthotrexate, azathioprine*) [...] sont discutés (The other **immunosuppressors treatments** (*mycophénolate mofétil, cyclophosphamide, méthotrexate, azathioprine*) [...] are discussed)

In NLP, two segments of text are considered paraphrases if similarity measures are high (such as cosine similarity or BLEU (Reiter, 2018), but these scores use only morphological or syntactic cues. Adversative paraphrases (with different lexical or syntactic forms, but with similar meaning) are more difficult to detect than paraphrases with few syntactic variations (Nighojkar and Licato, 2021). Paraphrase markers such as multi-word expressions (*c'est-à-dire* - 'that is to say', *signifie* - 'means', *est un/une* - 'is a') or punctuation signs, are often used to introduce paraphrases and they could help paraphrase automatic identification (Grabar and Hamon, 2015).

In our paper, we define **medical paraphrases** as different lexical representations that designate, simplify, or explain medical terms, while keeping a similar meaning (Fuchs, 2020; Vassiliadou, 2020; Buhnila, 2023). Our definition of the linguistic concept of paraphrase includes different types of word sequences, such as *definitions, rephrasing, exemplifications, explanations or abbreviations* (Eshkol-Taravella and Grabar, 2017; Buhnila, 2022b). We build a dataset of simple and multi-word terms linked to their **subsentential paraphrases**.The paraphrases could be simpler words or expressions, noun or verbal structures or simple enumerations of examples, often introduced by an explicit paraphrase marker. To illustrate our definition, we present some examples identified in our corpus. The term is displayed in **bold**, the paraphrase is in *italic* and the paraphrase marker that introduces the paraphrase is tagged with $< m >< /m >$:

- **distanciation physique** d'autres que cela <m>signifie</m> *couper les contacts sociaux* (**physical distancing** from others, which <m>means</m> *cutting off social contacts*);
- **l'anosmie**, <m>c'est-à-dire</m> *une perte totale de l'odorat* (**anosmia**, <m>meaning</m> *a total loss of sense of smell*).

We consider that medical paraphrases are useful for text simplification or adaptation. Simpler synonyms or hyperonyms might simplify the comprehension of the target audience, as well as definitions or exemplifications. Complex resources are required for such systems, but also various methods for producing them. Thus, we present related work on medical text simplification, paraphrase datasets or corpora and paraphrase identification or generation.

## 3. Related Work

**Text simplification in the medical domain** aims to explain or to replace scientific terms with simple words or paraphrases in order to enhance information accessibility to lay people (Grabar and Hamon, 2015, 2016; Cardon and Grabar, 2018; Koptient et al., 2019; Cardon and Grabar, 2021; Buhnila, 2022a). This simplified medical content might also be used to facilitate communication with patients (Pecout et al. 2019; Koptient and Grabar 2020). To simplify a medical text, two steps are necessary. Firstly, we identify medical terms, and secondly, we find the appropriate paraphrases for these terms. Both tasks are difficult. Automatic term identification based on terminological databases or ontologies with large coverage (such as **SNOMED** (Cote, 1998)) will not be able to identify newly created terms. For example, the Covid-19 pandemic created a large number of new terms, but they are not all included in the existing knowledge bases.[1] Tools for term

[1] After the end of this study, we came across a bilingual (French-English) ontology with Covid-19 terms accessi-

identification extract candidates from open-source texts and are more reliable, but the output has to be manually filtered (Rigouts Terryn et al., 2020).

**For the task of text simplification, paraphrase resources should relate terms to their paraphrases.** Most of the large paraphrase datasets contain sentential paraphrases from general language available in English: **MSRP** (*The Microsoft Research Paraphrase Corpus*) (Dolan et al., 2004), **PPDB** (*ParaPhrase DataBase*) (Ganitkevitch and Callison-Burch, 2014), **PAWS** (Zhang et al., 2019), (*Paraphrase Adversaries from Word Scrambling*). French language is represented in few resources (mostly multilingual), and only for the general domain, such as **PPDB**, **TaPaCo** (Scherrer, 2020) or **ParaCotta** (Aji et al., 2022). Subsentential paraphrases might be more appropriate to provide explanations or definitions for the terms, but few datasets containing subsentential paraphrases are available. One such resource is **PARADE** (He et al., 2020), a computer science dataset of definition-style paraphrases for English technical concepts extracted from online user-generated flashcards. These paraphrase datasets were built from general or computer science corpora, but they do not cover data from the field of medicine.

Due to the lack or medical paraphrase datasets or parallel corpora (original and paraphrases), NLP systems were developed for paraphrases identification or generation. Various statistical or deep learning methods were tested on paraphrase identification. Methods based on similarity measures, such as **Textual Semantic Similarity** (**STS**) (Agirre et al., 2016) or **Paraphrase Identification** (**PI**) (Brockett and Dolan 2005; Xu et al. 2015) identify paraphrases by counting words that have a certain degree of semantic equivalence and a similar lexical surface form. Various classifiers identify specific types of paraphrases based on syntactic criteria (Zhou et al., 2022). Sentence-level paraphrase identification methods are very effective for English datasets Peng et al. (2023) using BERT language model (Devlin et al., 2018). Again, few methods are designed to detect subsentential paraphrases. Linguistic patterns and n-grams are used to extract subsentential paraphrases from large medical comparable corpora (Cartoni and Deléger, 2011). Some methods use comparable corpora and **Abstract Meaning Representation (AMR)** (Bouamor et al., 2013) to detect subsentential paraphrases. These methods have some drawbacks when it comes to identify paraphrases with various surface forms for specific medical terms. Subsentential paraphrases, such as short definitions, exemplifications, explanations, or abbreviations might take different surface forms, but helps user's comprehension. Semantic similarity techniques fail to identify these types of paraphrases.

To avoid these drawbacks and to be able to create new paraphrase datasets, alternative methods, such as **Paraphrase Generation Method** (**PG**) (Gupta et al. 2018; Bowman et al. 2015) are employed to generate paraphrases with various forms, but similar meaning. Among these, the **APT** (*Adversarial Paraphrasing Task*) neural architecture (Nighojkar and Licato, 2021) uses a method for generating paraphrases with equivalent meanings and lexical and syntactic differences. This model identifies the general meaning of a sentence, not just the meaning of individual words. It is possible to infer the meaning from the term to the paraphrase and vice-versa.

In this paper, we present a dataset of subsentential paraphrases, as this type of paraphrase is not much exploited in the NLP community for the medical domain. In the next section, we present our project and our method used to create a large subsentential medical paraphrases dataset in French, LARGEMED. Moreover, we use this corpus as a resource for experimenting several methods for paraphrase classification and generation.

## 4. Method

The **ADAPTMED project** aims to create a large collection of terms and their paraphrases, by extending an existing subsentential paraphrase corpus **RefoMed** (Buhnila, 2023) with new terms from the Covid-19 pandemic and related topics such as social measures and vaccine campaign. Indeed, the Covid-19 pandemic generated a lot of new terms and paraphrases, frequently found in the Web archives created by the **National French Library (NFL)**[2].

We represent graphically our method in **Figure 1**. Firstly, to build a large paraphrase corpus, we identified the Web archives about the Covid-19 pandemic (a collection of Web pages dated from March 2020 to July 2020) maintained by the **NFL**. The archive contains a large number of new terms related to Covid-19, but also various paraphrases of this new terms, as people needed to better understand this new disease. The Web pages are available in several versions, due to frequent updates of the information during the pandemic. The pages are indexed with Apache Solr and the archives were manually explored with a specific query language. This query language is very complex and the requests had to be manually checked to identify the term and its paraphrase on the last version of the Web site. This step

ble here: https://www.hetop.eu/hetop/rep/fr/COVID/

[2]Bibliothèque Nationale de France (BNF)

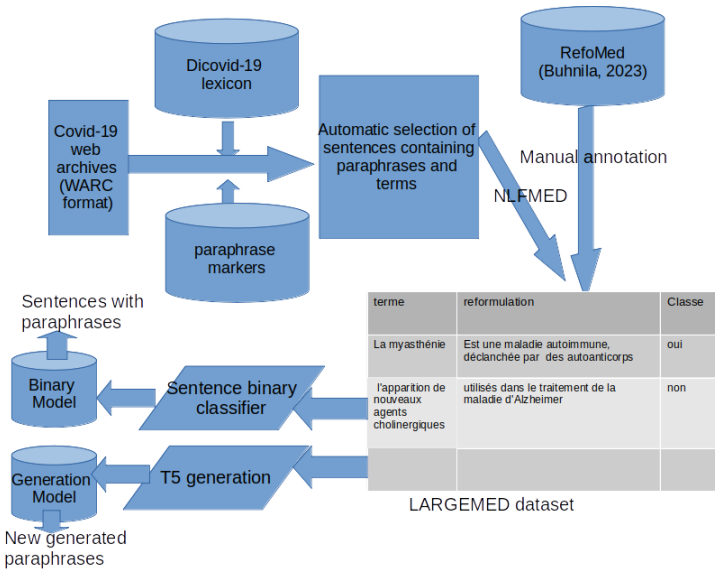


Figure 1: The method used to collect data and to develop the paraphrase classification and generation system.

is time-consuming, so the requests are send automatically to the Solr search engine to obtain a complete list of Web pages containing potentially terms and paraphrases, possibly accompanied by paraphrase markers.

From the Web pages which potentially contained paraphrases, we pre-selected the sentences that had both Covid-19 terms from the **Dicovid-19** dictionary [3] and paraphrase markers (expressions: *c'est à dire* - 'that is to say', *autrement dit* - 'in other terms'). These sentences were manually annotated and linguistically analysed. We present the annotation process in detail in section 5.3.2. The pre-selection of sentences containing paraphrases might reduce the number of sentences that should be manually annotated. Firstly, we set up a sentence binary classification using **CamemBERT** (Martin et al., 2019) to detect if the sentence contains a paraphrase in order to generalize the process for future experiments. Secondly, the paraphrases from our corpus are used to adapt a generation model **t5-base** (Raffel et al., 2020) for text generation to medical domain.

The next section presents the medical resources and the method of collecting subsentential paraphrases.

## 5. Resources

To compile and extend a large dataset containing terms and their subsentential paraphrases, we use a term list to select sentences containing terms, and therefore complete an existing paraphrase corpus RefoMed (Buhnila, 2023). We present this paraphrase corpus in the next section.

### 5.1. Existing Medical Paraphrase Corpus

**RefoMed** (Buhnila, 2023)[4] is a corpus of medical subsentential paraphrases in French and Romanian. The RefoMed corpus contains 11,653 pairs of medical terms and their medical paraphrases, 8,626 pairs in French and 3,027 pairs in Romanian. For this study we use only the French sub-corpus. The source corpora for French are **ClassYN** (Todirascu et al., 2012) and **CLEAR Cochrane** (Grabar and Cardon, 2018), both comparable corpora of scientific and simplified medical texts and abstracts. The RefoMed dataset was built by automatically extracting sentences that contain medical terms with the **SIFR-BioPortal** annotator (Tchechmedjiev et al., 2018), using the **SNOMED-3.5VF** medical ontology (Cote, 1998) (150,906 medical concepts). The sentences included in this corpus were selected if they contained both terms and paraphrase markers, such as *c'est-à-dire* ("so called"), *autrement dit*("in other words"), *également appelé*("also called"), *est une maladie*("is a disease"), *signifie* ("signifies / means") and punctuation signs, such as colons and brackets (Grabar and Hamon 2015; Antoine and Grabar 2016; Buhnila 2022b). The sentences were manually annotated and validated by 2 non-expert human coders. The coders follow the specific guidelines to annotate the status (if the sentence contains a paraphrase or not), the term, the paraphrase markers, the para-

[3] https://dicovid19.com

[4] https://github.com/ibuhnila/refomed

| Lang | Term | Mkr | Paraphase |
|---|---|---|---|
| **fr** | dyspnée | c'est-à-dire | gêne respiratoire |
| **en-tr** | dyspnea | i.e | difficulty breathing |

Table 1: An example of annotated paraphrase in French (**fr**) and its translation in English (**en-tr**). **Term** is the medical term automatically identified with Snomed, **Mkr** is the paraphrase marker that helps identifying the paraphrase, and **Paraphrase** represents a subsentential paraphrase.

phrase, the lexical relations (the paraphrase is a hyponym, hypernym or synonym to the term) and their semantico-pragmatic functions (definition, explanation, exemplification). The inter-coder agreement, computed as Krippendorff's $\alpha$ is moderate (0.61) for the paraphrase class. The validated term-marker-paraphrase pairs were included into **RefoMed**.
We build the new Covid-19 paraphrase corpus following the same method of selection of sentences containing a term from a lexical resource (Dicovid-19 in our case), and we follow the same annotation guidelines from (Buhnila, 2023), as explained in the section 5.3.2.

## 5.2. The Dicovid-19 dictionary

Several large coverage medical term databases are available, such as **UMLS** for English (Bodenreider, 2004) or **SNOMED International** for French (Cote, 1998), but they do not contain newly created terms related to the Covid-19 pandemic. Thus, we used the French **Dicovid-19** dictionary which contains 296 terms used or formed during the Covid-19 pandemic, such as *super spreader, vaccinodrome - 'mega vaccine center', N92 mask, distanciation sociale - 'social distancing', antivax - 'anti-vaccine collaborator'*. This dictionary is a key resource to select sentences containing Covid-19 terms and has been manually defined during the Covid-19 pandemic by a French lexicographer.

## 5.3. A new corpus NLF: Covid-19 Terms and Paraphrases

### 5.3.1. Data collection

The NLF Web archives contain 15TB of data and was build by automatic indexation of French Web pages such as newspapers, scientific blogs, popularisation blogs containing at least one mention of Covid-19 pandemic. Due to its size and the risk of incorrectly indexing web pages, functional words (such as punctuation, prepositions, conjunctions, simple verbs like *to be*, *to have*) were not included in the Solr search engine. Thus, we adapt our queries considering these constraints.
To collect the data we use expert queries including a term, a marker and a span window between them. Indeed, the query *text: "distanciation physique signifie" 7 AND (collections:"épidémie Covid-19")*, helps us to find the term **distanciation physique 'social distanciation'** along with the paraphrase marker **signifie - 'means'** (the number *7* indicates the word span). This query detected a paraphrase for the Covid-19 term *distanciation physique d'autres que cela signifie couper les contacts sociaux* (**physical distancing**: "physical distancing from others, which means cutting off social contacts"). The queries were manually written using Solr's interface.
Then, we manually selected Web pages and check if the page contained at least one Covid-19 term and its paraphrase in the same sentence. Afterwards, the url addresses were used to extract the text contained in the pages, by using the instance of the Apache Solr search engine.

**The next step was the semi-automatically extraction of sentences with term-paraphrase pairs**, introduced by paraphrase markers identified in the literature (Eshkol-Taravella and Grabar 2017; Buhnila 2022b). We asked the coders to identify the term, the paraphrase marker and the paraphrase as shown in **Table 1**.

We extracted 8,565 sentences containing at least a term and a paraphrase marker (out of 25,644 selected sentences). Through automatic annotation, we identified 893 pairs of terms and paraphrases in the same sentence (data is showed in **Table 2**). Only 10.42 % of sentences contained real paraphrases, manually validated. Additionally, we selected some definitions and paraphrases from Wikipedia Web pages of Covid-19 terms (140 sentences contain terms and their definitions or explanation). Then, we manually annotated them with lexical relations and semantic-pragmatic functions. We present the annotation process in the next section.

| Sent | Term | T-M Sent | C-Para | M-Para | Total Para |
|---|---|---|---|---|---|
| 25,644 | 8,565 | 1,725 | 637 | 176 | **893** |

Table 2: Quantitative data extracted from the url of the Covid-19 NLF archive collection. **T-M Sent** represents the number of sentences containing at least one term (**T**) and a marker (**M**); **C-Para** states the number of correct paraphrases (one per sentence); **M-Para** indicates the number of multiples paraphrases per sentence; **Total** represent the number of correct paraphrases.

#### 5.3.2. Annotation Process

To build the corpus, we follow the annotation method used for the RefoMed corpus (Buhnila, 2023). The RefoMed corpus was automatically annotated in terms and paraphrase markers and the paraphrases of medical terms were manually analysed from a lexical and semantico-pragmatic perspective following the guidelines provided by Eshkol-Taravella and Grabar (2017).

**Medical terms and paraphrase markers annotation.** Sentences containing both medical terms from the DiCovid-19 dictionary and the paraphrase markers are identified automatically using a rule-based method, applying regular expressions developed in Perl. Then, these sentences are manually annotated by at least two coders. The first task is to determine whether the sentences contain valid medical paraphrases or no paraphrase at all. Additionally, the term, the paraphrase marker and the paraphrase are also annotated.

**Lexical and semantico-pragmatic annotation.** The second task consists on the identification of lexical relations and the semantico-pragmatic functions of the paraphrases. On one hand, the lexical relations were defined as lexical links that exist between the two segments, the medical term and its paraphrase. These lexical relations can be synonymy, hypernymy, hyponymy and meronymy, as they are frequent in medical texts (Condamines 2018; Ramadier 2016; Săpoiu 2013). On the other hand, semantico-pragmatic functions represent the reasons that drives the speaker to use paraphrases in written medical texts, such as definition, rephrasing, designation, exemplification, or explanation (Eshkol-Taravella and Grabar 2017; Buhnila 2022b).

Thus, we obtained a new dataset, **NLFMED**, containing 1,033 medical paraphrases of Covid-19 terms, and a rich annotation following the same guidelines as for **RefoMed**. The two datasets are merged together into a larger dataset **LARGEMED** (17,393 sentences, annotated with terms, markers, paraphrases, lexical relations and semantic-pragmatic functions). This corpus is available for experiments of paraphrase classification and generation, in order to automatize data collection. These experiments are presented in the next section. Afterwards, we discuss the findings and limitations of our method for data collection.

## 6. Results and Discussion

Firstly, we evaluate the results of the annotation process applied in the **NLFMED** dataset. Secondly, we present the results from the classification and generation experiments conducted using this augmented paraphrase dataset.

### 6.1. Corpus Annotation and Evaluation

Only 1,725 sentences out 8,565 sentences containing Covid-19 terms contained both terms and paraphrase markers. To these sentences, we added 140 term definitions and explanations from the Wikipedia pages presenting the Covid terms. The annotation done by the two coders resulted in 1,033 correct paraphrases. We computed the Krippendorff's $\alpha$ score for several tasks: **a)** classification of sentences containing paraphrases; **b)** paraphrase markers; **c)** correct paraphrases; **d)** lexical relations and **e)** semantico-pragmatic functions.

For the task of sentence classification, we used the labels *"yes"* if the sentence contains a valid subsentential paraphrase and *"no"* - if the sentence contains no valid paraphrase. For this task, the inter-coder agreement is very high (**0,95**), meaning that the coders agreed in most of the cases. Then, we computed this agreement for the subsentential paraphrases : the coders agreed on recognizing a paraphrase in the sentence. The Krippendorff's $\alpha$ score was still very good (**0,80**) for this task as well as for the task of finding common discourse markers that introduce a paraphrase ($\alpha$=**0,82**). For the other elements that were annotated, the inter-coder agreement was good for the semantic-pragmatic functions ($\alpha$=**0,77**), but weaker for lexical relations ($\alpha$=**0,55**). Most cases of agreement concern the definition and the exemplification contexts, while paraphrases or explanations are more often subject of disagreement. For the lexical relation annotation, several confusions between meronymy and hyponymy or hyponymy/hypernymy (due to the reverse order of term and of the paraphrases) could be an explanation of a lower agreement score.

The existing **RefoMed** dataset and the newly built one from the Covid-19 Web archives NLF are compiled into a single medical subsentential paraphrase corpus for French **LARGEMED**. The same annotation guidelines are used to build both datasets. The method of building this corpus is mainly based on existing dictionaries (SNOMED for **RefoMed** and Dicovid-19 for the **NLFMED** dataset). If the terms are not found in the dictionary, then the sentences containing a paraphrase are not selected. In order to automate data collection, we conduct some experiments with the resulting corpus LARGEMED to build a binary model to detect if the sentence contains or not a paraphrase (section 6.2) or to adapt a generation model for creating variants of medical paraphrases (section 6.3). We present these experiments and the results obtained in the following subsections.

## 6.2. Binary Classification Experiments

The process of manual selection of sentences containing real paraphrases is time-consuming, but of high quality, when validated by human coders. In order to automatize the selection of sentences potentially containing paraphrases and to accelerate manual annotation, we built a binary classification model for detecting sentences containing paraphrases. For this purpose, we adapted the French **CamemBERT** language model (Martin et al., 2019) for the task of sentence classification, by pairing it with a set of 17,393 sentences manually annotated from the LARGEMED dataset. We used the information about paraphrase status (*yes* or *no*). We applied a cross-validation strategy with 5 and 10 folds, and we obtained the accuracy score of **0,84** and respectively **0,89**. From the several configurations of optimizers and loss functions, the *Adam* optimizer and the *SparseCategorialCrossentropy* loss function obtained the best results.
To compare this result with a bidirectional LSTM architecture, we use **CamemBERT** (Martin et al., 2019) to represent each sentence. The results show few variations between parameters such as the maximum length of the sentence containing or not paraphrases. However, we tried several configurations (embedding size of 150 and 200) and hyperparameters with the bidirectional LSTM architecture.

We obtained better accuracy results with CamemBERT when we used cross-validation (0,84, if we consider k=5 and 0,89 if we consider k=10) (see **Table 3**). For the bidirectional LSTM, we randomly selected 90% or 75% of the data for training, and we used several embedding size (150, 200). In this case, the accuracy was only 0,81.

| Train | Test | Embd size | Embd LM | Acc |
|---|---|---|---|---|
| 75% | 25% | 200 | $C$BERT | 0.81 |
| 90% | 10% | 150 | $C$BERT | 0.81 |
| Cross k=5 | - | 150 | $C$BERT | 0.84 |
| Cross k=10 | - | 200 | $C$BERT | 0.89 |

Table 3: Results of the classification task. **Train** represents the training split size, while **Test** is the test split size. **Embd size** is the **embeddings size** used for the experiments and **Emdb LM** represents the Language Model (LM) used for the task, which is the French LM CamemBERT ($C$BERT). For cross validation Cross, the values for k folds are available. We evaluate our results with accuracy (**Acc**).

We expected to obtain better result to automate the search of sentences with potential paraphrases. 11% of automatic annotation of the status of the sentences are errors, so this result should be improved. However, it is simpler to correct the automatic annotation rather than to do it from scratch. While we collected a large number of sentences from the Web archives, presumably containing terms and paraphrase markers, the sentence classifier helps reducing the time required to annotate the corpus, at least for the status task and will be useful to complete the dataset with new sentences containing potential paraphrases. For the other tasks, especially for lexical relation identification, the inter-coder agreement is too low to try to automatize the process.

## 6.3. Generation Experiments

As an alternative to data collection from existing Web pages, we propose to evaluate the quality of a paraphrase generation tool to obtain new paraphrases for the medical terms. Thus, we present the experiments using the new dataset LARGEMED in order to adapt a model to generate new medical paraphrases for the French Covid-19 terms. We adapt the APT neural architecture for adversative paraphrases and we use the T5 language model for generation and the dataset presented at section 5.3.

### 6.3.1. The APT Neural Network

The **APT** (*Adversarial Paraphrasing Task*) neural architecture (Nighojkar and Licato, 2021) uses a method for generating paraphrases with equivalent meanings but with lexical and syntactic differences at the surface level. This model identifies the general meaning of a sentence, not just the meaning of individual words. The APT architecture verifies if two sentences that are mutually implicit are also semantically equivalent. **APT** uses **BLEURT** (Sellam et al., 2020) to measure structure dissimilarity. **BLEURT** score evaluates automatically generated texts based on the word embeddings of the BERT language model (Devlin et al., 2018).
The corpus of paraphrases is used to adapt the APT paraphrase generation architecture for French medical data. **APT** generates paraphrases which have similar meanings (e.g. it is possible to infer the meaning of the term from the paraphrase and the term's meaning from the paraphrase).
The main changes of this strategy is the use of T5 model, available for French, which should be adapted for medical data, by using LARGEMED dataset including Covid-19 related terms and their paraphrases.

### 6.3.2. T5 Language Model

**T5** (*Text-to-Text Transformer*) (Raffel et al., 2020) was pre-trained on **C4** (*Colossal Clean Crawled Corpus*), a corpus with 7 terabytes of data extracted from the Common Crawl Web corpus. T5 had been trained for several specific NLP tasks, including

paraphrase identification and sentence similarity. We adapt it for our own dataset of subsentential medical paraphrases in French.

#### 6.3.3. Technical Aspects

We extract our experimental data from the **LARGEMED** paraphrases dataset (9,557 terms and their paraphrases from **RefoMed** and 1,033 paraphrases of Covid-19 terms from **NLF**). We fine-tuned the model t5_base with several configurations (the size of the paraphrase is 128 and 256 respectively): learning rate (3e-4), 4 epochs, the batch size (20), dropping rate (0,01), and AdamW optimiser (1e-8).

#### 6.3.4. Generation Results

We obtained 2,372 generated paraphrases for a test set of 576 terms contained in the test file (96 terms are related to Covid-19 pandemic). For each term, we obtained at most 5 paraphrase predictions. We analysed the predictions and annotated with *1* if the generated paraphrases are correct and *0* if they are incorrect.

| Predictions | Nb of terms | Percentage |
|---|---|---|
| At least 1 correct result | 204 | 35.41 % |
| No correct result | 372 | 64.59 % |
| Total | 576 | 100 % |

Table 4: The paraphrases generated (**Predictions**) by the T5 base model adapted for medical domain.

The paraphrases generated for 95 Covid-19 terms are generally quite far from the expected prediction. The few mentions of the Sars coronavirus or of the disease produce some paraphrases containing virus or disease with respiratory symptoms, but a large part of these terms do not generate valid output. We show some incorrect examples below, where *Truth* represents the initial paraphrase for the term, while *Prediction* represents the paraphrase generated by the language model.

- *Term*: **maladie à coronavirus 2019 (coronavirus disease 2019)**

  *Truth*: Covid-19 (Covid-19)

  *Prediction*: à transmission hépatique (hepatically transmitted)

- *Term*: **choc cytokinique (cytokine shock)**

  *Truth*: réponse exacerbée du système immunitaire inné (exacerbated response of the innate immune system)

  *Prediction*: une maladie de l'hémoglobine (a haemoglobin disease)

From all the predictions for the Covid-19 terms, we identify correct paraphrase predictions for 24 terms out of 96 from the Covid-19 term list. The correct paraphrases proposed are in general introduced by hypernyms: *Covid-19 longue* (long Covid-19) is paraphrased with *maladie chronique* (chronic disease); *la réplication virale* (the viral replication) is paraphrased with *une réplication de l'infection* (a replication of the infection).

We consider that the low performance of the language model in our experiments could be explained by the few occurrences of Covid terms in the training data set. Some Covid-19 terms design the measures to limit pandemic (*social distance, PCR test*) which are difficult to predict from the medical texts used to train the model. In the actual state of the model, few new paraphrases are provided if we compare with the paraphrases already available in the LARGEMED dataset.

## 7. Conclusion and Future Work

In this article we present a work in progress aiming to build a paraphrase corpus for medical terms collected from the Web archives of the National French Library and a method to extend this corpus by paraphrase classification and generation. Secondly, we follow the guidelines for annotating the paraphrases with lexical relations and semantico-pragmatic functions already applied for **RefoMed**. We created a new annotated resource of 1,033 Covid-19 related medical terms with their correspondent paraphrase **NLFMED** and compiled it into a larger French dataset **LARGEMED** (17,393 terms and their subsentential paraphrases). We obtained an accuracy score of 0.89 for the paraphrase classification task with CamemBERT. Still, it is possible to apply this classifier to pre-select sentences with paraphrases and then to refine by searching paraphrase markers and terms. The paraphrase generation is a difficult task. The results were not satisfactory for the Covid terms, due to the small size of our Covid-19 paraphrase dataset.

Future work includes enlarging the paraphrase Covid-19 dataset automatically with Solr extractions and then applying the binary classification to pre-select sentences containing paraphrases. Actually, the collection of new Web pages containing Dicovid terms is still in progress. The task of automatic paraphrase generation could give better results by combining APT with a language model adapted for the medical domain in French, such as CamemBERT-Bio (Martin et al., 2019) or DrBERT (Labrak et al., 2023), but also combining our dataset with other dataset available for general language. The final dataset will be used in a text simplification system for medical domain.

## 8. Acknowledgements

The ADAPTMED project has been funded by the National French Library (BNF) (https://www.bnf.fr) and supported by the National Library of the University of Strasbourg (BNU) (https://bnu.fr/fr), by FRLC (Research Network on Language and Communication) (https://frlc.hypotheses.org/) and the LiLPa research unit (University of Strasbourg) (https://lilpa.unistra.fr/).

## 9. References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.* ACL (Association for Computational Linguistics).

Alham Fikri Aji, Tirana Noor Fatyanosa, Radityo Eko Prasojo, Philip Arthur, Suci Fitriany, Salma Qonitah, Nadhifa Zulfa, Tomi Santoso, and Mahendra Data. 2022. Paracotta: Synthetic multilingual paraphrase corpora from the most diverse translation sample pair. *arXiv preprint arXiv:2205.04651*.

Edwide Antoine and Natalia Grabar. 2016. Exploitation de reformulations pour l'acquisition d'un vocabulaire expert/non expert. In *TALN 2016: Traitement Automatique des Langues Naturelles*.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270.

Houda Bouamor, Auréelien Max, and Anne Vilnat. 2013. Multitechnique paraphrase alignment: A contribution to pinpointing sub-sentential paraphrases. *ACM Trans. Intell. Syst. Technol.*, 4(3).

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Chris Brockett and William B Dolan. 2005. Support vector machines for paraphrase identification and corpus construction. In *Proceedings of the third international workshop on paraphrasing (IWP2005)*.

Ioana Buhnila. 2022a. Identifying medical paraphrases in scientific versus popularization texts in french for laypeople understanding. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 69–79.

Ioana Buhnila. 2022b. Le rôle des marqueurs et indicateurs dans l'analyse lexicale et sémantico-pragmatique de reformulations médicales. In *SHS Web of Conferences*, volume 138, page 10005. EDP Sciences.

Ioana Buhnila. 2023. *Une méthode automatique de construction de corpus de reformulation*. Ph.D. thesis, University of Strasbourg, France.

Rémi Cardon and Natalia Grabar. 2018. Identification of parallel sentences in comparable monolingual corpora from different registers. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 83–93.

Rémi Cardon and Natalia Grabar. 2021. Simplification automatique de textes biomédicaux en français: lorsque des données précises de petite taille aident. In *Traitement Automatique des Langues Naturelles*, pages 275–277. ATALA.

Bruno Cartoni and Louise Deléger. 2011. Découverte de patrons paraphrastiques en corpus comparable: une approche basée sur les n-grammes (extracting paraphrastic patterns comparable corpus: an approach based on n-grams). In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 182–187, Montpellier, France. ATALA.

Anne Condamines. 2018. Nouvelles perspectives pour la terminologie textuelle.

Josette Condamines, Anne ; Rebeyrolle. 1997. Point de vue en langue spécialisée. *Meta*, 42(1):174–184.

Roger A Cote. 1998. Systematized nomenclature of human and veterinary medicine: Snomed international. version 3.5. *Northfield, IL: College of American Pathologists*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

William Dolan, Chris Quirk, Chris Brockett, and Bill Dolan. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*.

Iris Eshkol-Taravella and Natalia Grabar. 2017. Taxonomy in reformulations from a corpus linguistics

perspective. *Syntaxe et semantique*, 18(1):149–184.

Catherine Fuchs. 1982. La paraphrase entre la langue et le discours. *Langue française*, La vulgarisation(53):22–33.

Catherine Fuchs. 2020. Paraphrase et reformulation: un chassé-croisé entre deux notions.

Juri Ganitkevitch and Chris Callison-Burch. 2014. The multilingual paraphrase database. In *LREC*, pages 4276–4283. Citeseer.

Natalia Grabar and Rémi Cardon. 2018. Clear-simple corpus for medical french. In *ATA*.

Natalia Grabar and Thierry Hamon. 2015. Extraction automatique de paraphrases grand public pour les termes médicaux. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 182–195.

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*, 57(1).

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. In *Proceedings of the aaai conference on artificial intelligence*, volume 32.

Yun He, Zhuoer Wang, Yin Zhang, Ruihong Huang, and James Caverlee. 2020. Parade: A new dataset for paraphrase identification requiring computer science domain knowledge. *arXiv preprint arXiv:2010.03725*.

Anaïs Koptient, Rémi Cardon, and Natalia Grabar. 2019. Simplification-induced transformations: typology and some characteristics. In *BioNLP 2019*.

Anaïs Koptient and Natalia Grabar. 2020. Fine-grained text simplification in french: steps towards a better grammaticality. In *International Symposium on Health Information Management Research*.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. Drbert: A robust pre-trained model in french for biomedical and clinical domains. *bioRxiv*, pages 2023–04.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

Animesh Nighojkar and John Licato. 2021. Improving paraphrase detection with the adversarial paraphrasing task. *arXiv preprint arXiv:2106.07691*.

Anaïs Pecout, Thi Mai Tran, 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*, 3(195):325–341.

Qiwei Peng, David Weir, and Julie Weeds. 2023. Testing paraphrase models on recognising sentence pairs at different degrees of semantic overlap. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 259–269, Toronto, Canada. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Lionel Ramadier. 2016. *Indexation et apprentissage de termes et de relations à partir de comptes rendus de radiologie*. Ph.D. thesis, Université Montpellier.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Ayla Rigouts Terryn, Véronique Hoste, Patrick Drouin, and Els Lefever. 2020. Termeval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (acter) dataset.

Camelia Săpoiu. 2013. *Hiponimia în terminologia medicală: modalităţi de abordare în semantică şi lexicografie*. Trend.

Yves Scherrer. 2020. Tapaco: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA).

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Andon Tchechmedjiev, Amine Abdaoui, Vincent Emonet, Stella Zevio, and Clement Jonquet. 2018. Sifr annotator: ontology-based semantic annotation of french biomedical text and clinical notes. *BMC bioinformatics*, 19:1–26.

Amalia Todirascu, Sebastian Padó, Jennifer Krisch, Max Kisselew, and Ulrich Heid. 2012. French and german corpora for audience-based text type classification. In *LREC*, volume 2012, pages 1591–1597.

Hélène Vassiliadou. 2020. Peut-on aborder la notion de" reformulation" autrement que par la typologie des marqueurs? pour une analyse sémasiologique et onomasiologique.

Wei Xu, Chris Callison-Burch, and William B Dolan. 2015. Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 1–11.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.

Chao Zhou, Cheng Qiu, and Daniel Acuna. 2022. Paraphrase identification with deep learning: A review of datasets and methods.