"The Concept of Sublanguage: A Comparative Study of a Professional Medical Corpus and a General Public Medical Corpus"

Pierson, Agathe

Abstract

This paper will present a comparative study, carried out to determine whether the medical language from the health section of Doctissimo's forum is a sublanguage in the same way as the medical language of professionals. We will firstly describe our three-step approach, and will introduce the typology that we developed and the tools used for this study. We will then proceed to evaluate the typology and analyse the syntactic and lexical-semantic observations that were carried out. Finally, we will review the current limitations of our methodology and will suggest some further research opportunities that our study opens up.

Document type : *Article de périodique (Journal article)*

Référence bibliographique

Pierson, Agathe. *The Concept of Sublanguage: A Comparative Study of a Professional Medical Corpus and a General Public Medical Corpus.* In: *Cahiers du CENTAL*, Vol. 8, p. 380 (2018)

The Concept of Sublanguage: A Comparative Study of a Professional Medical Corpus and a General Public Medical Corpus

Agathe Pierson

Université catholique de Louvain, Institut Langage et Communication, Cental

Abstract[1]

This paper will present a comparative study, carried out to determine whether the medical language from the health section of Doctissimo's forum is a sublanguage in the same way as the medical language of professionals. We will firstly describe our three-step approach, and will introduce the typology that we developed and the tools used for this study. We will then proceed to evaluate the typology and analyse the syntactic and lexical-semantic observations that were carried out. Finally, we will review the current limitations of our methodology and will suggest some further research opportunities that our study opens up.

Keywords: Sublanguage, medical language, causality, corpus linguistics

Résumé

Cet article présente une étude comparative effectuée pour déterminer si le langage médical provenant du forum Santé de Doctissimo est un sous-langage à l'égal du langage médical de professionnels. Tout d'abord, nous décrivons les trois étapes de notre approche et nous introduisons la typologie que nous avons élaborée ainsi que les outils utilisés pour cette étude. Ensuite, nous procédons à l'évaluation de la typologie et à l'analyse des observations syntaxiques et lexico-sémantiques que nous avons menées. Finalement, nous précisons les limites actuelles de notre méthodologie et nous suggérons quelques pistes de recherche que notre étude ouvre.

Mots-clés : Sous-langage, langage médical, causalité, linguistique de corpus

1. Introduction

Health and access to medical information embodies a significant social, economic and emotional issue for communities and individuals. With this in mind, it seems essential to study medical language in all its forms and variations, in order to access this information and understand the related issues. Thus, because of changes in production or media coverage, notably with the growth of the Internet, two categories of information coexist and overlap. These categories are general public information, produced by both non-specialists and experts who seek to popularise a specialised domain, and medical information, produced by specialists for specialists – transmitted by a specific language (Paganelli & Clavier, 2014: 2).

Both categories of information form a specific language known as 'sublanguage'. Harris (1971: 170-1) is the first to give a precise definition of this concept: "Certain proper subsets of the sentences of a language may be closed under some or all of the operations

[1] Acknowledgement: I would like to express my special thanks of gratitude to my colleague, Thomas François, for his help and comments on the statistical content. I would also thanks my PhD supervisor, Louise-Amélie Cougnon, for her support and suggestions.

defined in the language, and thus constitute a sublanguage of it.” A sublanguage is a specialised linguistic form that functions like a complete language and is described by a specific grammar. The grammar of a sublanguage may differ from that of a complete language in two ways. Firstly, a sublanguage grammar may not respect certain rules of complete language. Secondly, it may contain additional rules that are sufficiently recurrent and regular to be considered as representations of the specific structure of the sublanguage (Watrin, 2006: 127-128). In other words, a sublanguage is a subset of natural language, which meets lexical, syntactic and semantic restrictions of its own and which distinguish it from the general language. After Harris, Bross *et al.* (1972), Hirschman & Sager (1982) and Grishman & Kittredge (1986) revised the definition by paying particular attention to the context in which such sublanguage was used; a context which therefore corresponds to conditions of production, specific to a domain (e.g. the medical field), to circumstances (e.g. communication between experts) or to a particular theme (e.g. gastroenterology).

In this contribution, we would like to study the similarity between causal expressions and the vocabulary used in two medical corpora, one produced by health professionals and another produced by non-specialists, in order to determine whether non-specialists who talk about health use the same specialised language as professionals.

We will study the case of Doctissimo, the most visited French health website. Founded in 2000, the site has hosted over 8 million visitors per month since 2010, and has more than 40,000 articles related to health. Doctissimo also has a forum devoted to health that branches out into several categories (e.g. “health-child”, “diabetes”, “teeth”, etc.) (Sperlinga Gerner, 2015: 96). The forum allows contributors to seek advice, share experiences, give information and confirm a diagnosis.

Our basic assumption is that the medical language communicated by a non-specialist on the Doctissimo discussion forums (a sub-genre of Computer-Mediated Communication [Cougnon, 2012: 32]) takes on a professional nature, even though this form of medical language is not produced in a professional context. More precisely, we believe that this professional character would be translated into several linguistic levels, especially by the repetition of medical phraseology[2], as Battaïa (2016) argues about Net surfers. Indeed, as Battaïa points out (2016: 66-67):

> “The majority of Net surfers use the medical nomenclature, relying on technical terms [and constructions] to convey their comments. These contributions may be speeches from medical consultations or from the results of examinations and / or extracts from prescriptions.”

In other words, through this study, we will try to see how different production conditions exert an influence on the professional nature or the identification or differentiation of the language of the Doctissimo corpus as medical sublanguage.

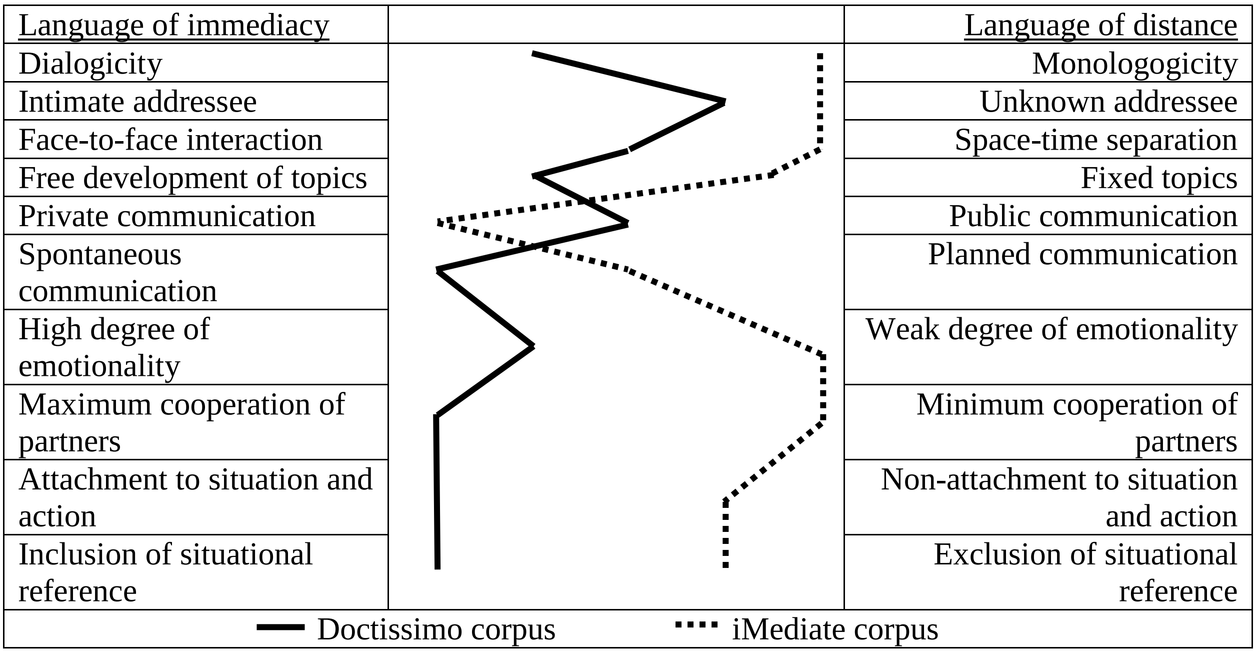
We will firstly present and compare the two corpora on which our study is based, as well as the methodology we developed to carry out this comparative study. We will also present a typology specifically designed for this analysis and the different tools used to annotate and make use of these corpora. We will then describe the initial results of the

[2] By *phraseology*, we mean “all the expressions (terminology and syntactic particularities) proper to a use, a milieu, an era, a writer.” (Our translation) (PR, 1993, 1935b, s.v. *phraséologie*)

study and discuss the problems attached to the notion of *causality* and those generated by our typology. Finally, we will detail the conclusions of this exploratory study and suggest potential avenues for further research.

2. Corpora and Methodology

In this section, we will describe the two corpora examined in the study and the methodology used to carry out the comparative analysis. In order to put these corpora into perspective, we will establish the communicative behaviour of the contributors in these two corpora.



Language of immediacy		Language of distance
Dialogicity		Monologogicity
Intimate addressee		Unknown addressee
Face-to-face interaction		Space-time separation
Free development of topics		Fixed topics
Private communication		Public communication
Spontaneous communication		Planned communication
High degree of emotionality		Weak degree of emotionality
Maximum cooperation of partners		Minimum cooperation of partners
Attachment to situation and action		Non-attachment to situation and action
Inclusion of situational reference		Exclusion of situational reference

Figure 1. Parameters characterising the communicative behaviour of interlocutors related to the situational and contextual determinants (adapted from Koch & Oesterreicher, 2001: 586)[3]

Figure 1 shows that the contributors of the iMediate corpus and those of the Doctissimo corpus differ radically in the way they express themselves. This is because of the kind of discourse they produce. In the first case a scientific text and, in the second, a spontaneous CMC production.

2.1. Corpora

2.1.1. Professional Medical Corpus (iMediate corpus)

The professional corpus is a digital corpus of around 225,000 medical texts (i.e. 93.4 million words), collected in six departments of a Brussels (Belgium) hospital (gastroenterology, MRI, bariatric and abdominal surgery, ultrasound and scanners). It consists of patient records from 1,000 patients over a period between 1996 and 2014. These texts are protocols for surgery, correspondence between doctors, letters to patients, visit notes, reports and examination sheets. This corpus is a product of the

[3] This figure is based on a five-degree continuum between the language of immediacy (left) to the language of distance (right). The degree of each parameter was established for both corpora on the basis of indices or observations made on the corpora.

iMediate project (Innoviris, 2014-2016) and is anonymised and secured for the needs of this project.

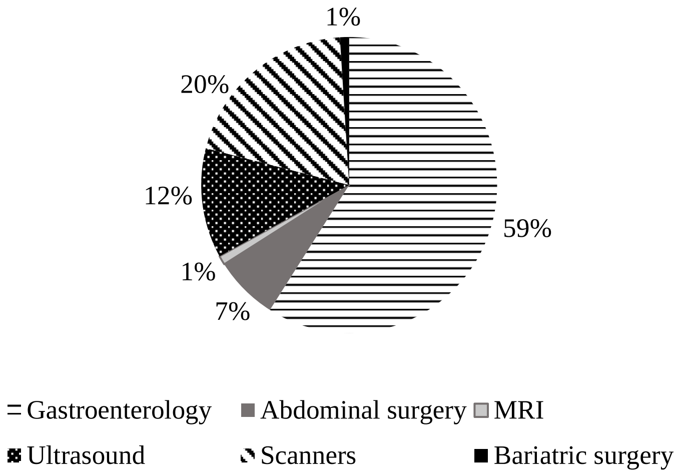


Figure 2. Contents of the iMediate corpus

2.1.2. Non-Professional Medical Corpus (Doctissimo corpus)

The non-professional corpus comes from the French discussion forum, Doctissimo. Although this corpus only includes the health section of the forum, it covers 97 different topics, roughly 920,000 threads, more than 23 million messages and more than 1.6 billion words, generated from 2000 to 2017. Of the 97 initial topics, we selected four (anorexia-bulimia, digestion-heartburn-gastric reflux, constipation-digestive transit and stomach ulcer), as they have a link with gastroenterology and are, therefore, more relevant for the lexical comparison. The corpus used for this study consists of about 76,000 threads, 1.8 million messages and more than 150 million words.

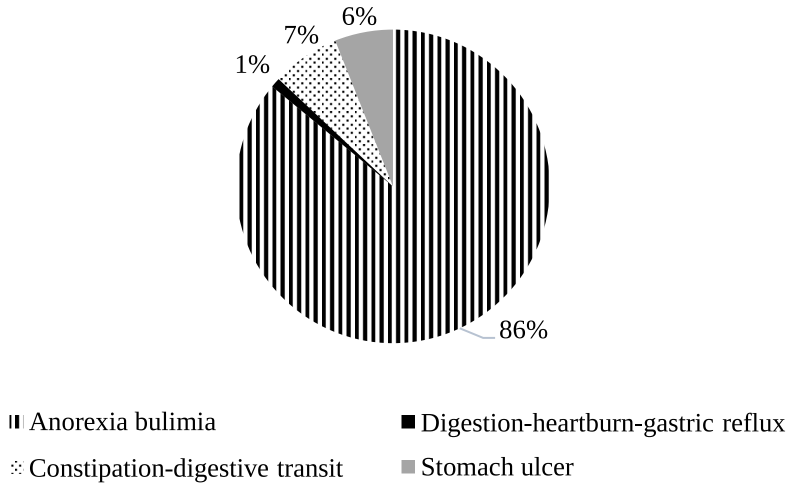


Figure 3. Contents of the Doctissimo corpus

2.2. Methodology

From the start, both corpora were evidently too large to be used directly for human annotation with the existing tools that are used to manually annotate and manage them (cf. sections 2.2.1. and 2.2.2.). Therefore, we decided to build a smaller representative sample of both corpora. We aimed to make these two samples conform as much as possible, in order to best respect corpus representativeness and to create a corpus that included the whole range of variability of the population, while respecting the proportion of each component of this population (Biber, 1993: 243-245).

2.2.1. Syntactic Study

As stated by Lehrberger (1982: 102), the high frequency of certain constructs (along with five other factors that he lists) play a role in the characterisation of a sublanguage. Causal expressions are generally widespread and have a variety of linguistic realisations (Nazarenko, 2000: 8-14). We hypothesise that there are causal expressions that are specific to medical sublanguage and that allow us to identify them as such.

For this syntactic study, we developed a new typology of causal expressions (*Appendix 1*). Previous work on causal typologies in French have been carried out (Jackiewicz, 1998; Nazarenko, 2000 and Gross, 2009), but for several reasons, these could not be used in their current form; although researchers noticed the existence of an implicit causality, they did not integrate it into their typology and offer no means to help in its identification. In addition, we found that there were more ambiguous causal expressions than those hitherto noted, such as concession conjunctions or ambiguous lexicon. Thanks to our exploratory analysis, we incorporated[4] new forms of causality that are ambiguous into this typology, as well as all the forms of implicit causality[5].

Two problems rapidly arose. On the one hand, by multiplying the types of causal expressions, it sometimes became complicated to associate just one single tag with an expression (example 1).

1) Antécédents médico-chirurgicaux appendicectomie vers 30 ans
Medico-surgical history of appendectomy around 30 years of age (Our translation)

The segment "antécédent médico-chirurgicaux" can be interpreted either as a title ("Medico-surgical antecedent: appendectomy towards 30 years of age", tagged by 'implicit_contextual_title'[6]), or as a lexical ambiguity, with the idea that it is potentially a cause, even if this causal interpretation is not the first meaning of this word. This segment would therefore be tagged by 'ambiguity_lexicon' (that is to say that the causal ambiguity concerns a notional word, such as a noun, verb or adjective, for which the primary interpretation is not causal, but which may have a causal interpretation in certain linguistic contexts). Indeed, in French, an antecedent is an "ailment prior to the currently considered illness" (TLF*i*, s.v. *antécédent*; our translation), without any mention of guaranteed causality. In order to circumvent this problem of overlap between categories,

[4] For examples that illustrate the typology, see *Appendix 2*.
[5] That is to say: every time a causal relation has to be inferred without any lexical item supporting it.
[6] This tag is used to indicate a causation that is not recognised by the use of a marker or a causal lexicon, but by its header position of the different sections of a file (e.g. *Appendix 2*).

the first principle we applied in the annotation process was to respect a hierarchy in the typology.

On the other hand, while we, as human annotators, can interpret the text and infer causal meanings where there is no explicit marker, it is almost impossible for a machine to locate such cases. Yet, one of the objectives of the general project in which this research is incorporated is to develop a programme that can automatically annotate causal expressions of French medical language. There are therefore still modifications and improvements to be made to this typology, as well as solutions to be found – such as identifying contexts conducive to the emergence of implicit causality – before developing such a programme.

We decided to use the web-based annotation tool WebAnno (Yimam, Eckart de Castilho, Gurevych & Biemann, 2014) in order to manually annotate the causal expressions[7]. The annotation process was carried out by one single annotator. We thus measured *intra*-annotator agreement (agreement of a single annotator over time).

2.2.2. Lexical-Semantic Study

The lexical-semantic analysis (the analysis of terminology and vocabulary) relies on two types of software with complementary performance that allow an optimal result for a first approach.

The first tool we used is *AntConc* (Anthony, 2014). This is a lexicometric analysis programme that enables the highlighting of the keywords of a corpus – those which stand out by virtue of their frequency – compared to a reference corpus. In this study, the reference corpus used to compare the two target corpora is a corpus combining the iMediate corpus and the Doctissimo corpus[8]. Therefore, comparing the two samples to this reference corpus will make it possible to highlight lexical items that appear more frequently in one sample compared to the other. The software also displays the numbers of types and tokens per file: this allows for the quick calculation of the types-tokens ratio (TTR), indicating the level of lexical variation of a text.

The second tool we used is *Tropes* (Molette & Landré, 1994), a programme used for semantic analysis, which identifies the key elements and the key events that make up the structure of a text. Therefore, Tropes makes it possible to uncover the named entities of a text, the relationships that link them and the hierarchy of these relationships.

[7] This decision was made based on a comparison of existing linguistic annotation tools (Fort, 2012). This tool makes it easy to calculate inter-annotator-agreement, according to several standard metrics depending on the type of data or annotation (Cohen's Kappa for categorical data; Fleiss' Kappa for more than two annotators and Krippendorff's Alpha, preferentially for continuous data). We also chose WebAnno because it allows the export of annotated segments with their tags, and positions them in a clean, machine-processable format (.tsv).

[8] Hence, the definition of the reference corpus differs, for this study, from the commonly accepted definition: “A reference corpus is one that is designed to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference materials.” (Sinclair, 1996: 10)

3. Results

This study sought to examine the extent to which the medical language used on the Doctissimos's discussion forum is similar to medical language, a sublanguage, whose semantic, syntactic and lexical constraints have already been the subject of numerous research projects (Friedman, Lyman & Sager, 1987; Ceusters et al., 1998; Fleischman, 2005; Grouin et al., 2011; Denecke, 2014).

3.1. Assessment of the Typology

One of the most valuable ways to determine the effectiveness of a typology is to calculate the inter- or the intra-annotator agreement. While numerous agreement metrics exist, it is generally accepted that adequate inter-annotator scoring accounts for agreement that occurs by chance (Fort & Claveau, 2012: 386-387). Calculating this figure amounts to wondering to what extent the typology proposed for the annotation task is sufficiently precise and objective, so as to leave little room for subjectivity, interpretation and coincidence. WebAnno automatically calculated Cohen's Alpha (given that the data is categorical). The table in Figure 4 highlights the general type of annotated causal expression (explicit, ambiguous or implicit), then the position of the item and the item itself. The last two columns detail examples of the labels assigned in two different annotation rounds. During the first annotation session, we annotated 585 causal expressions in the iMediate corpus and 1,034 in the Doctissimo corpus; during the second annotation session, we identified 591 causal expressions in the iMediate corpus and 1,067 in the Doctissimo corpus. At the end of our annotation campaign and by combining the two annotations, we obtain a total of 665 annotations for the iMediate corpus and 1,284 for the Doctissimo corpus.

General Type	Position of the annotation	Annotator 1	Annotator 2
Explicit	1302-1305 [tel]	Exp_Conj_Indir_Csq	Exp_Conj_Indir_Csq
Ambiguity	12428-12443 [dans le sens où]	Ambig_Conj_Contex	Exp_Conj_Dir
Implicit	1615-1616 [,]	Imp_Para	Null

Figure 4. Examples of the three possible cases of inter-annotation

According to Santos' scale (2015: 2), the Cohen's Alpha scores are very high (0.98 for the Doctissimo corpus, 0.96 for the iMediate corpus, see Figure 5). One reason for such a high score is that the annotator is identical in both annotations and is still largely in agreement with herself[9], even after an interval of three weeks. This high score is further explained by the fact that, with WebAnno, only exact span matches[10] were taken into account in the calculation of agreement.

[9] Although the annotator often seems to agree with herself, some examples that contradict this impression have been identified. Indeed, since this is a discussion forum, the Doctissimo corpus has the particularity of doubling some extracts when these are quoted by a user who wants to refer to themselves. Sometimes, the same annotator annotates the original message differently and takes it back into another message.

[10] We use the expression *exact span match* when the annotation of the first annotator exactly matches with the annotation of the other annotator, and vice versa; the expression *loose span match* is used when

	WebAnno	Editing
Doctissimo corpus	0.98	0.58
iMediate corpus	0.96	0.74

Figure 5. Intra-annotator agreement (WebAnno) – Corrected intra-annotator agreement

In order to verify the calculation of agreement, we compared the figures provided by WebAnno with R (R Development Core Team, 2008): we calculated agreement on exact span matches and loose span matches. Agreement figures are presented in Figure 4. This difference of intra-annotator agreement between the two corpora may be understood in light of the observation that causality is easier to detect in the language of specialists, which adheres more to standards and codes of writing than the language of non-specialists. This is true even of the implicit, which often lies precisely in the headings of the sections of the reports (e.g. *Concerne* [Concern] and *Indication* [indication]) or in the line breaks. Indeed, in medical language between non-specialists, spontaneity and unpredictability in the expressions – including causal ones – reign (Battaïa, 2016: 55).

Gross (2009: 1-12) argues that causality is a complex notion to be detected because the boundaries of its markers are not easily identifiable. This explains the difference between the intra-annotator agreement scores calculated by WebAnno and those computed by R. In this way, when an explicit lexical marker of causality is tagged, annotators tend to assign the same type (example 2). However, when annotating an implicit marker, annotators are more likely to disagree on the type (example 3).

2) […] et un travail que j'adore que je considère comme mon deuxième bébé **puisque** c'est ma propre boîte.

   […] and a work that I love and that I regard as my second baby **because** this is my own company. (Our translation, we emphasise)

3) Je n'ai pas de carence**,** je mange normalement aux autres repas de la journée.

   I do not have a deficiency**,** I eat normally in the other meals of the day. (Our translation, we emphasise)

An annotator annotating on two different occasions may not distinguish the same causal relationships within a text and, even if a particular causal relationship is recognised, they may not annotate it in the exact same position. For instance, of the 1,284 causal expressions annotated in the Doctissimo corpus, 800 received the same tag at the same position and 19 are recognised in exactly the same place but with different tags.

the expression annotated by the first annotator only partially matches with the expression annotated by the other annotator, and vice versa. We use the term *disagreement* when a single annotator has annotated an expression.

Consequently, there are 465 discrepancies in the annotations, which are either due to the same causal relationship not being retained in the two annotation processes, or to the fact that it was picked out both times but not in the same place.

Furthermore, as well as the rule of hierarchy (section 2.2.1.), further rules could be applied in order to:

- determine whether a relationship is causal or not (e.g. as it provides an answer to the question "Why?" [Nazarenko, 2000: p.15]). The sentences could be rephrased using the verb *to cause* or by introducing *because/then*, etc.;
- add a confidence score to each annotation, which would present the degree of certainty of the annotator – added by the annotator in order to indicate how sure they are about their decision.

3.2. Syntactic Analysis

The first syntactic observation formulated here is related to the proportion of causality in the two annotated corpora. We noticed that there are nearly twice as many causal expressions in the Doctissimo corpus (1,284 annotated relations for 55,929 words, or 2.3%[11] of the total corpus) than in the iMediate corpus (664 annotated relations for 54,620 words, or 1.2% of the total corpus). This result is surprising because we intuitively expected to see more causal expressions in the professional language corpus. It seemed obvious to us that every physician would generate a large number of causal expressions in order to justify the conclusions of a particular diagnosis or medical procedure to their colleagues and the patient. We also hypothesise that the rate of causal expressions is twice as high in the Doctissimo corpus, partly because it contains two text types relying enormously on causal expressions. Indeed, discussion forums include scientific information (Battaïa, 2016: 69), which uses both causal expressions, because it must be objective and valid in the same way professional medical language is, and testimonies, in which the participants are asked to explain the reasons for their shared experience and to discuss their concerns and medical situation. Causality is even more important in the expression of everyday narratives, because it helps to dispel uncertainties, to understand the links between various events or states, to structure the universe around us, and to understand what is happening to us (Gross, 2009: 1). It is therefore logical to find more causal expressions in the Doctissimo corpus than in the iMediate corpus.

To further investigate possible explanations for this first observation, we will compute a significance test[12], the independent chi-squared test (see Figure 6). The variable (*corpus type*) appears to influence ($\chi^2_{(2)} = 27.068$, $p < 0.001$) the distribution of causal

[11] We express ourselves as a percentage, even if the two units are not identical (causal expressions that can contain zero or more words as opposed to words); the two rates remain well comparable.

[12] All statistical computations for this part are carried out using the SPSS statistical software (IBM, 2013).

expressions in the generic categories (*explicit*, *ambiguous*, and *implicit*). In other words, our corpus types (medical language written by specialists in the form of scientific reports and medical language transmitted via discussion forums by non-specialists) significantly influence the causal expressions produced, at the level of the three generic categories. However, this dependence relation is weak, as shown by Cramer's V of 0.118 (p <0.001), which means that the type of corpus, and therefore the language it conveys, has only little influence on the use of causal expressions. If we take V^2, which roughly corresponds to the proportion of the variance of the type of causal expressions explained by the type of corpus, it amounts only to 1%. It is clear that the corpus type – which combines both the diatopic variation and the speaker's degree of specialisation – does not alone explain the distribution of the causal categories. The variability within types of causal expressions must therefore also be influenced by other factors.

	Supra-categories of causal expressions			
Type of corpus	Explicit	Ambiguous	Implicit	Total
Doctissimo	419 (32.6%)	379 (29.5%)	486 (37.9%)	1284 (100%)
iMediate	172 (25.9%)	159 (23.9%)	333 (50.2%)	664 (100%)
χ^2	27.068 (p=0.001)			
Cramer’s V	0.118 (p=0.001)			

Figure 6. Chi-squared and Cramer’s V for the supra-categories of causal expressions

These results are opposed to our initial hypothesis: the medical language of the iMediate corpus does not contain more explicit causal expressions than the other types of expressions. These results could be explained by the fact that medicine is not an exact science, as it does not make claims of absolute truth. Consequently, although health professionals will frequently resort to causal expressions, they will be subtler and less assertive. This trend perhaps reflects a desire not to assert without certainty and to maintain a certain reserve with respect to observations.

It is perhaps not surprising that the implicit constitutes the largest proportion of causal expressions in the Doctissimo corpus, given that:

> “It is in conversation[13] that the enunciator can be particularly “economical”. It is in conversation that he can count most on the cooperation of the recipient and make full use of the possibilities of immediate regulation of communication, in the event of a malfunction. […] It is in live verbal exchanges that para- and non-verbal factors of communication intervene to provide mutual comprehension.” (Dispy, 2011: 67, our translation)

We also examined what occurs at the subcategory level (tags), in order to expand on our existing results. Hence, we computed the independent chi-squared test to determine whether the variable (*corpus type*) influences the distribution of causal expressions in the various categories identified in the typology. Our calculation confirmed ($\chi^2_{(22)}$ = 661.917, p <0.001) that the type of corpus significantly influences the type of causal

[13] Discussion forums constitute a fully-fledged conversation, even if it may experience delayed or sustained exchanges.

expressions at the level of the sub-categories. This time, we can describe this dependence relation as (relatively) strong, as proved by Cramer's V of 0.561 (p <0.001). This figure shows that the language conveyed by each of these two corpora has a moderately strong influence on the types of causal expressions performed. By taking the V^2 into account, we find that 31.5% of the variance in type of causal expression may be explained as a result of the type of corpus. According to these results, it is evident that at a more advanced level, the "corpus" variable accounts for almost one third of the distribution of the different causal categories, although other factors – thus far unknown – are to be taken into account to explain this variability.

In conclusion, we must further extend our study to the level of each tag, and even to the level of each concrete expression. Indeed, perhaps *parce que* and *car*, both direct causal conjunctions in French, do not have the same distribution in both corpora because each fulfils a distinct function (e.g. real cause vs. cause of utterance, link between two syntactic phrases vs. link between two clauses, etc.). What is absolutely essential to this analysis is that, whatever the level of study chosen, the type of corpus influences the causal expressions that will be produced. Consequently, the distribution functions of these expressions are not identical on these two corpora. We therefore come to the conclusion that these two corpora operate differently, as far as syntax is concerned, and they do not currently result from the same language.

We also pursued several lexical-semantic analyses. We initially looked into lexical richness. We believe that comparing the lexical richness of the Doctissimo corpus and the iMediate corpus may bring to light a tendency towards similarity or dissimilarity in the lexical profiles of these two corpora. To quantify this richness, we measured the type/token[14] ratio; that is, the number of types (the number of distinct words in a corpus) over the number of tokens (the total number of words in a corpus, regardless of how often they are repeated). We observed that there is no notable difference between these two ratios (Figure 7). This shows a similarity between the two corpora, which have a relatively low lexical richness. This weakness is explained by the fact that the two corpora correspond to textual styles in which lexical research, the desire to distinguish oneself by one's literary style, does not take precedence. Indeed, in both corpora, the essential function is the intended function, whether it takes the form of a follow-up (iMediate), a request or advice (Doctissimo), rather than the form under which the message is transmitted.

	Tokens	Types	TTR	Std TTR
Doctissimo	2858	22040	0.13	13%
iMediate	2773	13672	0.2	20%
Global	5631	35712	0.16	16%

[14] The lexical units are segmented by AntConc according to white spaces and punctuation (*delimiter* characters), except the dash.

Figure 7. Main lexicometric characteristics of the 2 corpora of medical language

At the level of the lexicon, we generated lists of keywords and kept the words with more than 15 occurrences[15]: 104 units for iMediate and 170 units for Doctissimo. First of all, the greater number of units in the Doctissimo corpus reinforces the finding of a greater lexical richness in the iMediate corpus, as it shows that the speakers in the iMediate corpus use more lower-frequency words than those in the Doctissimo corpus, who have a stronger tendency to reuse the same words. Moreover, we found that the rates of grammatical words (determiners, prepositions and pronouns) and of function words (conjunctions) are higher in the Doctissimo corpus (35%) than in the iMediate corpus (24%). This difference in ratios is not surprising, as it can be easily explained by the importance of the phatic function[16], which is almost omnipresent in this communication medium, where contact with others is primordial and the will to transmit the right message and fairest emotion is a priority. Another explanation of this phenomenon could be that the contributors in the Doctissimo corpus feel more attached to the situation and action, and therefore include more situational reference.

As for the use of pronouns, we notice a clear difference between the two corpora (Figure 8):

	Doctissimo corpus	iMediate corpus
Je (*I*)	65.2%	31.4%
Tu (*You*)	15.9%	3.4%
Il/Elle (*He/She*)	7.5%	19.1%
Vous (*You*)	1.6%	20.1%

Figure 8. Use of pronouns by corpus[17]

We observe that, where the contributors of the Doctissimo corpus express themselves to a large extent in the first person with *je* (*I*) and use the pronoun of informal address *tu* (informal *you*), those of the iMediate corpus, even if they also use *je (I)*, are more distant towards their interlocutor (*vous* – formal *you*) and are more oriented towards the outside, especially towards their patient, than towards their own practice or experience. This may be explained by the fact that the Doctissimo corpus partly consists of personal testimonies, in which the subject retraces their own experience.

[15] We restricted our corpus to the terms with more than 15 occurrences in order to have sufficiently frequent words to be representative. For certain results, we also looked at *hapax* – terms that have only one occurrence.

[16] The phatic function is one of the six functions of language listed by Jakobson (1985: 114). It is the function used by a contributor to establish, maintain or interrupt physical or psychic contact with the addressee. For example, the phrase *Tu vois?* (*You know?*) is the concern of the phatic function.

[17] We only used the pronouns which were of interest for our comparative study in this table.

At the verb level, we observed that more verbs are used in the Doctissimo corpus than in the iMediate corpus, and that these do not distribute similarly as, of the words with more than 15 occurrences, only 6% are verbs in the iMediate corpus, compared to 15% in the Doctissimo corpus. Moreover, we note that in the Doctissimo corpus, the majority of verbs are conjugated in the simple present, which is one of the characteristic features of CMC (Cougnon, 2012: 57). This tense feature should usually be interpreted as a historical (narrative) present. The contributors of the iMediate corpus favour variation, using as many past forms, which express either a sequence of actions or a completed action, as present forms, which exercise the same function of historical present, partaking in the description of events and medical acts. However, they use fewer future and conditional forms, which mainly transmit an order, instruction or advice.

Type of verbs	Examples of verbs	Doctissimo corpus	iMediate corpus
Verbs expressing actions (*verbes factifs*)	To exclude, to explore, to chain up	52.1%	40.8%
Verbs expressing states or notions of possession (*verbes statifs*)	To be, to remain, to persist, to match	26%	30.4%
Verbs expressing a declaration about a state, an action, a being, an object, a feeling (*verbes déclaratifs*)	To believe, to envisage, to suggest, to misjudge	17.8%	28.1%
Verbs expressing an act by and in language (*verbes performatifs*)	To admit, to assure	4.1%	0.7%

Figure 9. Types of verbs by corpus according to Tropes

We find that, proportionally, the iMediate corpus contains more verbs expressing actions than the Doctissimo corpus, which, itself, has more verbs expressing states or declarations about a feeling (Figure 9). In practice, this is manifested through different lexical verbs. Where health professionals use action verbs such as *effectuer*, *mesurer*, *controller*, *scanner*, *opérer* (*to carry out*, *to measure*, *to monitor*, *to scan*, *to operate*) – verbs that describe medical acts, treatments and procedures – the non-professionals prefer emotional verbs[18], which appeal to the emotions of the readers (*aimer, espérer, adorer, s'inquiéter* [*to love*, *to hope*, *to adore*, *to worry*]) and verbs that reflect the behaviour they adopt in the situation described (*manger moins, boire plus, cuisiner, profiter* [*to eat less, to drink more, to cook, to enjoy*]).

[18] *Emotional verbs* must be distinguished from *verbs expressing a feeling* insofar as emotional verbs describe emotions or express psychological activities without the proposition having any real or false value. The verbs expressing a feeling are propositional verbs that express the mental state of the subject on a certain proposition, which may be either true or false (Bittar, 2010: 25).

Finally, we turn to the question of nouns and adjectives. These are the most represented categories in the iMediate corpus, as 27% and 15% of the corpus words are, respectively, nouns and adjectives, whereas they only correspond to 12% and 5% of the Doctissimo corpus. These results reflect the fact that the two corpora do not contain the same language. Specialist languages, conversely, display fewer connectors and moderating particles. Now, by comparing our two corpora, we observe that the iMediate corpus contains 7.5% of these grammatical words, whereas the Doctissimo corpus contains 16%. Indeed, as Watrin (2006: 127) points out, medical language "contains a very large majority of telegraphic [and elliptic] syntactic structures", which lead to the disappearance of articles and prepositions and to the use of incomplete verbal structures due to the suppression of auxiliary or modal verbs. This increases the apparent dichotomy between the two corpora.

We are able to highlight even more trends: the most frequent nouns of the iMediate corpus are generally related to the medical profession (*examen, traitement, confrère, blessure, prescripteur, acte, patient* [*examination, treatment, fellow member, lesion, prescribing doctor, act, patient*]) or to the event described (*date, genre, jour* [*date, gender, day*]). The rarer they are[19], the more specific they become, thus requiring medical knowledge to use them appropriately (*hémostase, ictérique, stigmata, nodule* [*haemostasis, icteric, stigma, nodule*]). At any point in the text, the health specialists refer to their practice, use precise medical terminology, and even incorporate specific terms in a desire for rigor and accuracy.

By comparison, the contributors of the Doctissimo corpus use nouns related more directly to everyday life, to their reality (*repas, soir, matin, plaisir* [*meal, evening, morning, pleasure*]), their diet and their gastric problems (*ventre, poids, protéine, lait,* [*belly, weight, protein, milk*]), or even generic nouns such as *chose* (*thing*). Although we can observe that they are also concerned about their health and seek to share related information, we note, however, two elements that tend to prove that they do not use specialised terminology as much as professionals. Firstly, when they might be led to use a medical term, they opt for a common version[20]. Secondly, contrary to what we observe in the iMediate corpus, the rarest words are not the most specific ones; they simply belong to other areas of daily life (*ambiance, internet, marketing, gaz, restaurant, cinéma [atmosphere, internet, marketing, gas, restaurant, cinema]*). Hence, while they also speak of health and illness, the contributors of the Doctissimo corpus do so using

[19] We used the frequency, as calculated by Tropes, over the entire corpus.

[20] For instance, they say "J'arrive à remanger des choses bon pour mon corps et qui apporte des choses" ("I manage to eat things that are good for my body and that give things") when they could have said "Je cuisine de la nourriture saine qui apporte des nutriments" ("I cook healthy food that gives nutrients"); they say "Je ne sais pas ce qu'il faut que je boive pour échapper à mes problèmes d'estomac" ("I do not know what I need to drink to avoid stomach problems") when they could have used a specialised term "Je ne sais pas quelle substance cholalogue soulagerait mon reflux acide" ("I do not know which cholalogic substance would ease my acid reflux").

an accessible and understandable vocabulary that reduces the level of expertise but increases the sharing, the exchanges and the feeling of connection with the community.

This was made all the more evident when we analysed the “Reference Universes”[21].

	iMediate corpus	Doctissimo corpus
Macrotheme	Health (911 semes)	Diet (410 semes)
	Body (323 semes)	Health (144 semes)
Microtheme	Disease (390 semes)	Food (402 semes)
	Medicine (318 semes)	Time (199 semes)
	Digestive system (94 semes)	Family (66 semes)
	Organ of secretion (66 semes)	Weight (41 semes)
	Drug (50 semes)	Disease (38 semes)
	Organ (43 semes)	Cuisine (38 semes)

Figure 10. Reference Universes detected by Tropes by frequency of semes by corpus

It is evident from this table (Figure 10) that even if the two corpora deal with the same subject (health), one does so consistently with an outward and detached approach towards the situation and uses precise terms that recall specific domains and ontologies; the other approaches the subject of health more introspectively, whereby the speaker is really living what they are discussing and linking these remarks with other parameters of life and reality that influence this experience. This last point is also supported by the types of adjectives present in each corpus. Indeed, Tropes lists three classes of adjectives: objective adjectives (which indicate the existence or absence of a property), subjective adjectives (which indicate a value judgment or an emotional response) and numerical adjectives (which group numbers and numerical adjectives). For the iMediate corpus, 50.4% of the adjectives were tagged as objective (*fébrile, bulbaire, incomplete* [*febrile*, *bulbar*, *incomplete*],) compared to 24.7% and 24.9% as subjective and numeric adjectives, whereas for the Doctissimo corpus, 47.3% of the adjectives were characterised as subjective (*à fleur de peau, miraculeux, tristounet* [*sensitive, miraculous, sorrowful*]), compared to 27.1% and 25.5% as objective and numeric adjectives.

4. Conclusion

As demonstrated by our analysis, the iMediate corpus, written by specialists in the form of medical reports, and the Doctissimo medical corpus, involving non-specialists in a hybrid communication medium, do not have the same degree of specialty at the

[21] These Reference Universes correspond to semantic fields that are detected by Tropes on the basis of a dictionary of semantic equivalents.

linguistic level. It is clear that the professionalism and skill of the contributors of the iMediate corpus leads to what does indeed constitute a sublanguage; this is less clear for the Doctissimo corpus. Contrary to our expectations, and on the scale of our study, it appears that the language of Doctissimo cannot be accorded the definition of sublanguage.

As previously mentioned, causal expressions present a different distribution, which is at least partially explained by the corpus in which they appear. In addition, it was demonstrated by lexical-semantic analyses that the vocabulary and the semantic fields used in the Doctissimo corpus are not homologous to those involved in the iMediate corpus. The latter is firmly fixed in the medical field, whereas the former also refers to the dimension of everyday life. Consequently, if the medical language that appears in the iMediate corpus belongs to a sublanguage, as it meets the requirements of this notion, the Doctissimo corpus cannot be defined in the same way, as it does not adopt the same linguistic functions.

This demonstrates, all things considered, that the circumstances in which a language is produced and the profile of the speaker greatly influence the specialisation and degree of expertise emerging from a discourse.

In order to support these conclusions, it would be interesting to analyse a larger quantity of data, once an effective automatic causal annotation programme has been developed. Before we could at arrive that point, further annotations would be needed to assess the accuracy and coverage of the typology we are proposing. More research is also required to broaden the comparison to additional types of medical language. This could be achieved by adding the medical language used by health professionals intending to popularise medical information (on health websites or by interventions on discussion forums) and the medical language used by non-professionals who demonstrate important medical knowledge due to their familiarity with certain chronic diseases (serious diabetes, cystic fibrosis, multiple sclerosis) to our current corpora. Eventually, since a non-specialist of the medical world usually refers to several physicians – notably by reporting indirect speech (von Münchow, 2004), it would be relevant to study the extent to which these reported discourses actually translate the medical language as used by medical professionals.

References

Anthony L. (2014), *AntConc (Version 3.4.3) [Computer Software]*. Waseda University, Tokyo, Japan. Available at http://www.laurenceanthony.net/.

Battaïa C. (2016), “Information médicale et émotion dans les forums de santé”, in Colloc J. and Hénocque B. (eds), Enjeux du big data et identifications des données médicales, Lavoisier: Cachan, vol. 12: 51-72 (Les cahiers du numérique).

Biber D. (1993), “Representativeness in Corpus Design”, in *Literary and Linguistic Computing*, vol. 8/4: 243-257.

Bittar A. (2010), *Building a TimeBank for French: A Reference Corpus Annotated According to the ISO-TimeML Standard*, PhD Thesis, Prof. Danlos L., Amsili P. and Denis P., 19 November 2006, Université Paris Diderot.

Bross I.D.J., Shapiro P.A. and Anderson B.B. (1972), “How information is carried in scientific sub-languages”, in: *Science*, vol. 176: 1303-1307.

Ceusters W., Spyns P., De Moor G. and Martin W. (1998), *Syntactic-Semantic Tagging of Medical Texts: The Multitale Project*, IOS Press, Amsterdam.

Cougnon, L.-A. (2012). *L'écrit sms: variations lexicale et syntaxique en francophonie*. PhD Thesis, Prof. Fairon C., 23 September 2012, Université catholique de Louvain.

Denecke K. (2014), “Sublanguage analysis of Medical Weblogs”, in: Lovis C., Séroussi B. and Hasman A. (eds.), *E-Healt – For continuity of care. Proceedings of MIE2014*, IOS Press, Amsterdam: 565-569.

Dispy M. (2011), *Pour étayer l’apprentissage de l’implicite*, Presses universitaires de Namur, Namur (Tactiques, 6).

Fleischman S. (2005), “Language and Medicine”, in: Schiffrin D., Tannen D. and Hamilton H.E. (eds), *The Handbook of Discourse Analysis*, Blackwell Publishers Ltd, Malden (Massachusetts): 470-502.

Fort K. (2012), *Les ressources annotées, un enjeu pour l’analyse de contenu: vers une méthodologie de l’annotation manuelle de corpus.* PhD Thesis, Prof. Nazarenko, A., 7 December 2012, Université Paris 13.

Fort K. and Claveau V. (2012), “Annotation manuelle de matchs de foot: Oh la la la! l’accord inter-annotateurs! et c’est le but!”, in: Antoniadis G., Blanchon H. and Sérasset G. (eds.), *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*: vol. 2, ATALA and AFCP, Grenoble: 383-390.

Friedman C., Lyman M.S. and Sager N. (1987), *Medical Language Processing: Computer Management of Narrative Data*, Addison-Wesley, MA).

Grishman R. and Kittredge R. (1986), *Analyzing language in restricted domains*, Hillsdale, New Jersey.

Gross G. (2009), *Sémantique de la cause*. Peeters, Leuven (Collection linguistique, 96).

Grouin C., Deléger L., Cartoni B., Rosset S. and Zweigenbaum P. (2011), “Accès au contenu sémantique en langage de spécialité: extraction des prescriptions et concepts médicaux”, in: *Proceedings of TALN 2011 (Traitement automatique des langues naturelles*), Montpellier: 109-120.

Harris Z.S. (1971), *Structures mathématiques du langage*. Dunod, Paris.

Hirschman L. and Sager N. (1982), “Automatic Information Formatting of a Medical Sublanguage”, in: Kittredge R. and Lehrberger J. (eds.), *Sublanguage. Study of Language in Restricted Semantic Domains,* Walter de Gruyter, Berlin – New-York: 27-80.

IBM Corp. (2013), *IBM SPSS Statistics for Windows, Version 22.0.*, IBM Corp, Armonk (NY).

Imbs, P. (ed.). (1971-1994), *Trésor de la langue française. Dictionnaire de la langue du XIXe et du XXe siècle (1789-1960)*, Gallimard, Paris. 16 vol. En ligne: http://www.cnrtl.fr/definition (TLFi).

Innoviris (2014-2016) *iMediate: Interoperability of Medical Data through Information extraction and Term Encoding.* Project supervised by C. Fairon.

Jackiewicz A. (1998), *L’expression de la causalité dans les textes. Contribution au filtrage sémantique par une méthode informatique d’exploration contextuelle*, PhD Thesis, Prof. Desclés J.-P., 21 February 1998, Université Paris-Sorbonne.

Jakobson R. (1985). **“**Metalanguage as a linguistic problem”, in: Rudy S. (ed.), *Selected Writings, VII. Contribution to Comparative Mythology*, De Gruyter Mouton, Berlin, 113-121.

Koch P. and Oesterreicher W. (2001), **“**Gesprochene Sprache und geschriebene Sprache/Langage parlé et langage écrit**”**, in: Holtus G., Metzeltin M. and Schmitt C. (eds.), *Lexikonder Romanistischen Linguistik,* vol: 1/2, Niemeyer, Tübingen: 584–627. Cited in: Cougnon L.-A. and Bouraoui J.-L. (2017), “Orality and Scripturality of Telephony and SMS”, in *Romance Languages in the Media*.

Lehrberger R. (1982), “Automatic Translation and the Concept of Sublanguage”, in: Kittredge R. and Lehrberger J. (eds.), *Sublanguage. Study of Language in Restricted Semantic Domains,* Walter de Gruyter, Berlin – New-York: 81-106.

Nazarenko A. (2000), *La cause et son expression en français*. Ophrys, Gap (L’essentiel français).

Paganelli C. and Clavier V. (2014), “S'informer via des médias sociaux de santé : quelle place pour les experts ?”, in *Le Temps des médias*, vol. 23/2: 141-143.

R Development Core Team (2008), *R: A language and environment for statistical computing. R Foundation for Statistical Computing,* Vienna. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rey A. and Rey-Debove J. (eds.) (1993), *Le nouveau petit Robert 1 : dictionnaire alphabétique et analogique de la langue française*. Dictionnaires Le Robert, Paris (PR).

Santos F. (2015), *Le kappa de Cohen: un outil de mesure de l’accord inter-juges sur des caractères qualitatifs,* CNRS, UMR 5199 PACEA.

Siepmann D., Bürgel C. and Diwersy S. (2016), “Le *Corpus de référence du français contemporain* (CRFC), un corpus massif du français largement diversifié par genres”, in *5^e Congrès Mondial de linguistique française*, n° 27.

Sinclair, J. (1996), *Preliminary Recommendations on Corpus Typology*. EAGLES Document EAG-TCWG-CTYP/P.

Sperlinga Gerner M.-M. (2015), *Variations graphiques des textes des forums sur Internet*, PhD Thesis, Prof. Pellat J.-C., 30 September 2015, Université de Strasbourg.

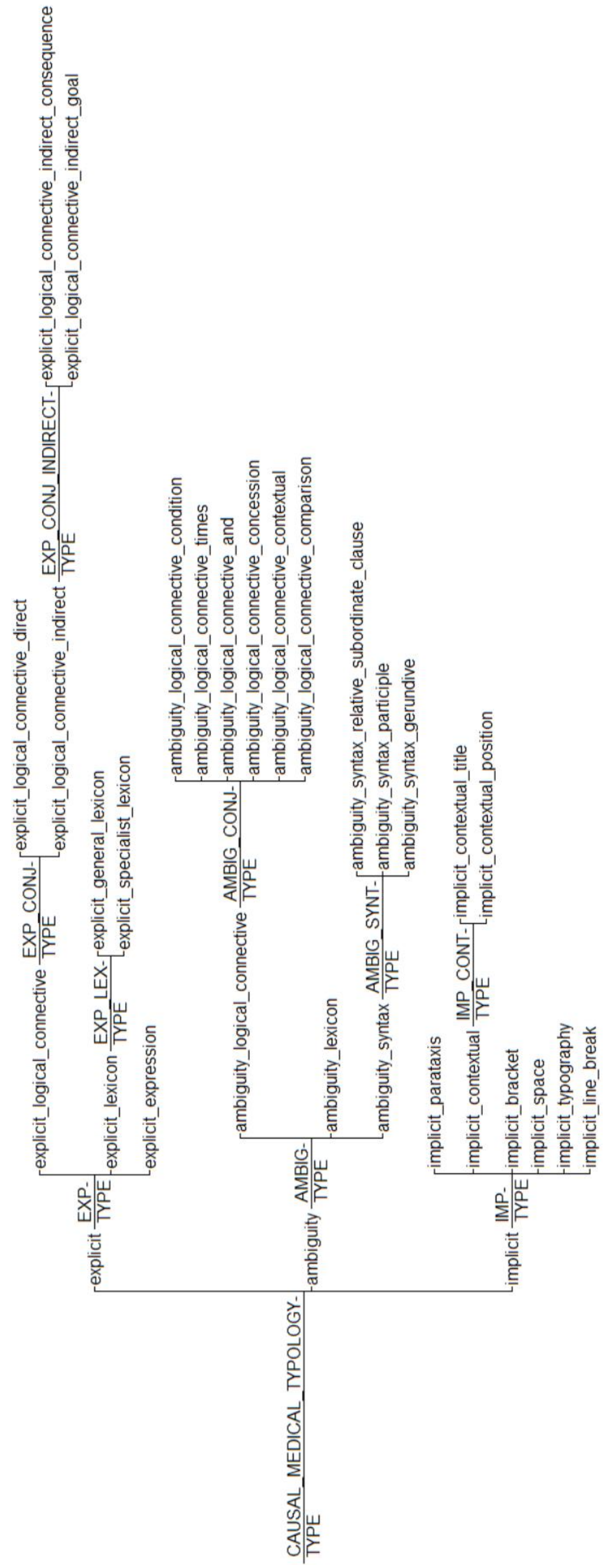
Tropes VF 8.4 (2014), Software application created by Molette P. and Landré A., based on the work of Ghiglione R., Université Paris VIII.

von Münchow P. (2004), “Le discours rapporté dans un forum de discussion sur l’internet”, in: *Les Carnets du Cediscor*, vol. 8: 91-111.

Watrin P. (2006), *Une approche hybride de l’extraction d’information: sous-langages et lexique-grammaire*, PhD Thesis, Prof. Fairon C., 26 October 2006, Université catholique de Louvain.

Yimam S.M., Eckart de Castilho R., Gurevych I. and Biemann C. (2014), “Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno”, in *Proceedings of ACL-2014, demo session (Baltimore, MD, USA).*

Appendix 1. Typology of the causality


CAUSAL_MEDICAL_TYPOLOGY-TYPE
explicit
ambiguity
implicit
EXP-TYPE
explicit_logical_connective
explicit_lexicon
explicit_expression
EXP_CONJ-TYPE
explicit_logical_connective_direct
explicit_logical_connective_indirect
EXP_CONJ_INDIRECT-TYPE
explicit_logical_connective_indirect_consequence
explicit_logical_connective_indirect_goal
EXP_LEX-TYPE
explicit_general_lexicon
explicit_specialist_lexicon
AMBIG-TYPE
ambiguity_logical_connective
ambiguity_lexicon
ambiguity_syntax
AMBIG_CONJ-TYPE
ambiguity_logical_connective_condition
ambiguity_logical_connective_times
ambiguity_logical_connective_and
ambiguity_logical_connective_concession
ambiguity_logical_connective_contextual
ambiguity_logical_connective_comparison
AMBIG_SYNT-TYPE
ambiguity_syntax_relative_subordinate_clause
ambiguity_syntax_participle
ambiguity_syntax_gerundive
IMP-TYPE
implicit_parataxis
implicit_contextual
implicit_bracket
implicit_space
implicit_typography
implicit_line_break
IMP_CONT-TYPE
implicit_contextual_title
implicit_contextual_position

Appendix 2. Illustrations of the different types of causal expressions

Explicit

Explicit_logicial_connective_direct

Fr : « Aucune CI au traitement nécessaire **car** pas d'angiome […] »

En: “No contraindication to treatment required **because** no angioma”

Explicit_logical_connective_indirect_consequence

Fr : « Le fond de la plaie […] a **donc** également été réséqué. »

En: “**So** the bottom of the wound was also excised.”

Explicite_logical_connective_indirect_goal

Fr : « […] on réalise également une acquisition tardive **afin de** bien visualiser les uretères. »

En: “A late acquisition is also carried out **in order to** well visualise the ureters.”

Explicit_general_lexicon

Fr : « Un kyste sacro-coccygien […] **est responsable d'**un suintement persistant. »

En: “A sacro-coccygeal cyst **is responsible for** a persistent oozing.”

Explicit_specialist_lexicon

Fr : « […] ulcère bulbaire hémorragique aigu, probablement **iatrogène**. »

En: “Sharp haemorrhagic bulbar ulcer, probably **iatrogenic**.”

Explicit_expression

Fr : « La morphologie du patient et la gravité de la péritonite **rendent** l'abord laparoscopique **impossible**. »

En: “The patient’s morphology and the seriousness of the peritonitis **make** the laparoscopic approach **impossible**.”

Ambiguity

Ambiguity_logical_connective_condition

Fr : « **Dans l'hypothèse d'**une éventuelle diarrhée sanglante d'origine infectieuse […] »

En: “**In the event of** a possible bloody diarrhea of infectious origin”

Ambiguity_logical_connective_times

Fr : « **Après** injection, on réalise également une acquisition tardive […] »

En: “**After** injection, a late acquisition is also carried out”

Ambiguity_logical_connective_and

Fr : « […] je mangeais rien de la journée **et** le soir je me goinfr[ai] de cochonnerie en tout genre. »

En: “I ate nothing of the day **and** in the evening I get stuffed with junk in any kind”

Ambiguity_logical_connective_concession

Fr : « La plaie, **bien que** large et profonde, est très propre […] »

En: “The wound, **although** wide and deep, is very clean.”

Ambiguity_logical_connective_contextual

Fr : « Ces crises surviennent à chaque fois **dans** un contexte assez stressant. »

En: “These crises happen each time **in** a rather stressful context.”

Ambiguity_logical_connective_comparison

Fr : « […] **plus** tu varies ton alimentation, **moins** t'as des problèmes de transit ! »

En: “**The more** you change your diet, **the less** you have digestive transit problems!”

Ambiguity_lexicon

Fr : « […] élément qui **suggère** une irritation du péritoine. »

En: “element that **suggests** irritation of the peritoneum.”

Ambiguity_syntax_relative_subordinate_clause

Fr : « [...] l'eau est retenue par mon corps **qui** a peur d'en manquer. »

En: “The water is restrained by my body **which** is afraid of missing it.”

Ambiguity_syntax_participle

Fr : « […] étant fille unique, je comprends parfaitement ce que tu ressens. »

En: “Being single girl, I understand perfectly what you feel.”

Ambiguity_syntax_gerundive

Fr : « […] c'est **en se restreignant** que ce sera pire. »

En: “It is **by restricting** that it will be worse.”

Implicit

Implicit_parataxis

Fr : « Je ne connais personne dans ma situation**,** je ne sais pas avec qui en parler […] »

En: “I do not know anyone in my situation**,** I do not know with whom to talk about it.”

Implicit_contextual_title

Fr : « **Données cliniques :** Dolichocolon spastique connu »

En: “**Clinical data:** Known spastic dolichocolon”

Implicit_contextual_position

Fr : « Oesogastroduodénoscopie » (at the very beginning of the file)

En: “Oesogastroduodenoscopy”

Implicit_bracket

Fr : « 1 tasse de café avec du lait et deux sucre **(**j'arrive pas à boire le café sans sucre […]) »

En: “1 cup of coffee with milk and two sugar **(**I can not drink sugar-free coffee […])”

Implicit_space

Fr : « […] je ne suis pas un modèle pour les repas je prends des 0% […] »

En: “I'm not an example for meals I take 0%”

Implicit_typography

Fr : « […] je sens que j'ai quand même du chemin à faire **:** ça me semble ÉNORME ce que je mange […] »

En: “I feel I still have a way to go**:** it seems HUGE what I eat”

Implicit_line_break

Fr : « L'anuscopie révèle un état inflammatoire de la muqueuse.

Attitude proposée »

En: “Anoscopy reveals an inflammatory state of the mucosa.

Proposed attitude.”