L'expression de la causalité dans le langage médical. Établissement d'une typologie spécifique

Agathe Pierson
Université catholique de Louvain

Abstract
La cause revêt une importance particulière dans le langage médical. En effet, l'apparition de symptômes ou encore la prise de décision médicale efficace sont exprimées par des relations causales. Il semble donc primordial de réaliser une étude linguistique des expressions causales en vue de caractériser les spécificités du langage médical relatives à cette relation discursive et de permettre ainsi une meilleure compréhension de la structure de l'information dans le domaine médical. Dans cette contribution, nous présentons les premiers résultats d'une étude sur l'expression de la causalité dans le langage médical, étude effectuée afin de développer des programmes d'extraction et d'analyse de la causalité. Premièrement, nous revenons brièvement sur la notion de *sous-langage*, qui sous-tend toute notre étude. Ensuite, nous décrivons les trois corpus ainsi que le logiciel d'annotation que nous avons utilisés. Après quoi, nous développons notre méthodologie de la causalité appliquée au langage médical en l'illustrant d'exemples authentiques. Nous procédons alors à l'évaluation de cette typologie dont nous soulignons les limites et pour laquelle nous soumettons des pistes d'amélioration. Finalement, nous synthétisons les découvertes actuelles de notre recherche et nous suggérons quelques pistes de recherche que notre étude ouvre.

1. Introduction

Dans cet article, nous présentons la typologie de la causalité que nous avons établie pour analyser les expressions causales et la distribution de celles-ci dans le langage médical dans le but de discriminer le sous-langage médical du langage ordinaire et d'en analyser les spécificités d'usage et de fonctionnement.

La première étape de notre étude des expressions causales, dont nous proposons de rendre compte dans cette contribution, consiste en l'élaboration d'une typologie de la causalité linguistique dont la granularité doit répondre aux besoins et aux objectifs de notre enquête, c'est-à-dire être adaptée au français écrit et au langage médical. À terme, nous espérons pouvoir développer un système d'extraction automatique qui réponde à ces diverses contraintes et qui vienne ainsi combler quelques-unes des lacunes précédemment relevées.

Cette étude repose sur le constat que « l'une des principales préoccupations de tous les départements médicaux est la causalité », notamment parce que les « docteurs diagnostiquent leurs patients sur la base de leurs symptômes et de leur histoire » (Kleinberg & Hripcsak 2011: 1102). Des progrès dans les approches linguistico-computationnelles de la causalité pourraient déboucher sur un impact majeur dans tout ce qui touche aux décisions cliniques supportant la santé publique. En effet, la relation causale est d'une importance particulière en médecine, qui est concernée par l'apparition de symptômes et de signes médicaux, par le développement de traitement et de médicaments ou encore par l'identification d'effets secondaires, ces différents phénomènes étant linguistiquement exprimés au moyen d'expressions causales (Khoo, Chan & Niu 2000, Morlane-Hondère, Grouin, Moriceau & Zweigenbaum 2015).

Il semble primordial de réaliser une étude linguistique des expressions causales en vue de caractériser les spécificités du langage médical relatives à cette relation discursive et de permettre

ainsi des systèmes informatiques plus performants. S'il existe déjà des outils d'extraction des expressions causales dans le langage médical (Itto & Bouma 2011, Sorgente, Vettigli & Mele 2013), ceux-ci répondent de manière limitée à certaines nécessités entrainées par les objectifs de notre recherche. Ainsi, ces outils n'intègrent que peu de techniques de désambiguïsation (Khoo, Chan & Niu 2000) et de prise en compte de l'implicite (Girju & Moldovan 2002). De même, peu d'études jusqu'ici réalisées portent sur la langue française (Garcia 1997, Corminboeuf 2010). Finalement, si ces outils extraient effectivement les expressions ou relations causales, leurs développeurs ne produisent pas d'analyse linguistique visant à comprendre les phénomènes linguistiques qu'ils extraient. Lorsque nous aurons développé des programmes d'extraction performants et répondant aux spécificités de notre corpus, nous aimerions proposer un examen linguistique approfondi de ces observations, notamment en étudiant les fonctions propres à chaque type d'expression causale, dépendant de variables telles que le service d'hospitalisation ou le lieu d'émission du document, et en confrontant ces analyses aux professionnels en réalisant une enquête sociolinguistique par questionnaire.

Ce papier est composé de quatre parties. Dans la première partie, nous reprenons succinctement la théorie relative au concept de *sous-langage*. Dans la deuxième partie, nous présentons les différents corpus sur lesquels se base cette étude, ainsi que les outils utilisés. Dans la troisième partie, nous développons la typologie. Enfin, dans la quatrième et dernière partie, nous commentons et évaluons provisoirement cette typologie en livrant quelques statistiques descriptives.

2. La notion de sous-langage

La notion de sous-langage est au cœur de notre contribution puisque nous partons du postulat que le langage médical est un sous-langage qui fonctionne comme un langage complet et bénéficie d'une

grammaire qui lui est spécifique, ce qui se traduirait notamment par un usage et un fonctionnement particuliers des expressions causales.

Le concept de *sous-langage* a été pour la première fois défini par Harris (1971: 170-1) : « Certains sous-ensembles propres de phrases d'un langage peuvent être fermés pour certaines (ou toutes les opérations définies dans le langage), et constituer ainsi un sous-langage de ce langage. » De la définition harrissienne émergent deux potentialités : d'une part, il serait possible de définir des règles précises relatives à la grammaire de ce sous-langage ; d'autre part, il serait possible de compiler un corpus qui renfermerait la totalité du lexique et de la syntaxe de ce sous-langage. D'autres auteurs, tels que Bross *et al.* (1972), Hirschman & Sager (1982) ainsi que Grishman & Kittredge (1986), ont revu et augmenté cette première définition pour inclure une attention particulière au contexte dans lequel un sous-langage était utilisé, contexte qui répond à des conditions de production spécifiques à un domaine particulier, à certaines circonstances qu'à un sujet limité. Ainsi, le sous-langage touche à trois dimensions essentielles de la communication : le lexique, la sémantique et la syntaxe, chacune connaissant des restrictions propres au sous-langage (Watrin 2006: 127-128).

Lehrberger (1982: 102) ajoute à ces premiers critères de reconnaissance d'un sous-langage ceux de (i) la haute fréquence de certaines constructions ; (ii) une structure textuelle qui repose sur des sections caractéristiques du sous-langage et sur un ordonnancement des phrases qui prend un sens spécifique dans ce sous-langage ; (iii) l'utilisation de symboles spéciaux. Le critère de récurrence de certaines structures peut, par exemple, s'appliquer aux expressions causales, qui sont largement répandues et qui connaissent une grande variété dans leurs réalisations linguistiques (Nazarenko 2000: 8-14). Le développement de Lehrberger nous permet de poser l'hypothèse qu'il existe un emploi spécifique des expressions causales dans le domaine médical, spécificité qui nous autoriserait à l'identifier plus nettement encore comme un sous-langage.

3. Méthodologie

Dans cette section, nous présentons le corpus d'étude principal, à savoir un corpus médical rédigé par des professionnels du domaine, ainsi que deux corpus de langue par rapport auxquels nous élaborons des comparaisons.

Quand on cherche à caractériser un type de langage, il s'agit de comparer ce langage à d'autres formes de langage. Nous avons choisi de réaliser parallèlement deux études dans le but d'offrir le plus de couverture possible au phénomène causal et, partant, une meilleure compréhension. D'une part, nous comparons les usages de notre corpus d'intérêt à ceux d'un corpus de langue générale[1] afin de montrer en quoi notre sous-langage s'en distingue et afin de confirmer qu'il existe une utilisation particulière au langage médical utilisé entre professionnels. D'autre part, nous confrontons deux corpus partageant la même thématique[2], mais se distinguant par l'expertise médicale de leurs auteurs et par le médium de communication, afin de rattacher l'un ou l'autre corpus à la notion de sous-langage.

3.1 Corpus d'étude

3.1.1 Corpus médicaux

Nous commençons par caractériser les deux corpus médicaux en les mettant en perspective sur le plan du comportement communicatif de leurs locuteurs. En effet, si nous voulons comparer ces deux corpus, il nous semble essentiel de garder en tête les variables qui peuvent influencer les conclusions de nos comparaisons.

---

[1] C'est-à-dire que ce corpus a la volonté de « représenter » la totalité de la langue française dans tous ces usages et par tous ces canaux.

[2] Étant présupposé que le corpus iMediate, présenté dans la section 2.1.1.1, correspond à la définition harrissienne de sous-langage et répond donc à ses restrictions ; là où les corpus Doctissimo et de référence ne semblent *a priori* pas cadrer avec cette définition harrissienne.

La figure 1 montre que les contributeurs des deux corpus médicaux sur lesquels se fondent nos analyses, le corpus iMediate et le corpus Doctissimo, diffèrent dans leur façon de s'exprimer. Ceci tient sans doute au type de discours qu'ils produisent : dans le premier cas un texte de spécialiste en contexte professionnel et, dans le second, une production spontanée de CMO (Communication Médiée par Ordinateur, Cougnon 2012: 57).

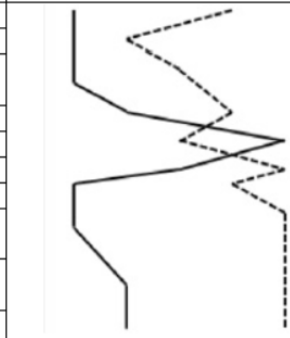
Langue de la distance		Langue de la proximité
Monologue		Dialogue
Interlocuteur inconnu		Interlocuteur intime
Séparation spatio-temporelle		Coprésente spatio-temporelle
Fixation thématique		Liberté thématique
Communication publique		Communication privée
Communication préparée		Communication spontanée
Émotionnalité faible		Émotionnalité forte
Coopération communicative minime		Coopération communicative intense
Détachement actionnel et situationnel		Ancrage actionnel et situationnel
Détachement référentiel de la situation		Ancrage référentiel dans la situation
— Corpus iMediate ····· Corpus Doctissimo		

Figure 1. Paramètres caractérisant le profil communicatif des locuteurs en fonction des facteurs situationnels et contextuels déterminants (Koch & Oesterreicher, 2001 : 586)[3]

3.1.1.1 Corpus médical rédigé par des professionnels (Corpus iMediate)

Le corpus du langage médical professionnel est un corpus informatisé de 226 133 textes médicaux (88 millions de mots) collectés dans six services d'un hôpital bruxellois (*cf.* Figure 2). Il est constitué de 1 000 fichiers-patients (1 fichier correspondant à un patient) hospitalisés entre 1996 et 2014. Ce corpus réunit différents sous-genres de textes médicaux, tels que des protocoles opératoires,

[3] Ce graphe est construit au départ d'un continuum de cinq degrés entre la langue de la distance (à gauche) et la langue de la proximité (à droite). La situation des corpus par rapport à chacun des paramètres fut établie sur la base des indices observés sur les corpus.

du courrier entre médecins ou au patient, des suivis de consultation, etc. Il est le produit du projet iMediate (Innoviris 2014-2016) et est anonymisé ainsi que sécurisé pour les besoins de notre projet.

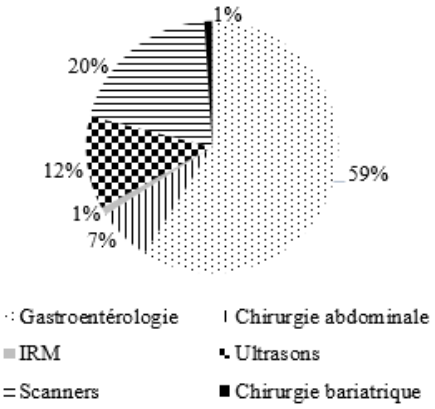


Figure 2. Constitution du corpus médical rédigé par des professionnels (Corpus iMediate)

3.1.1.2 Corpus médical rédigé par des non-professionnels (Corpus Doctissimo)

Le corpus de non-professionnels provient du site de service proposant des forums de discussion en français *Doctissimo* ; il s'agit du site français relatif à la santé le plus visité (Sperlinga Gerner 2015: 96). Nous n'avons retenu que la partie du site consacrée aux forums dédiés à la santé (plus de 40.000 articles), parmi d'autres thèmes tels que la psychologie, la nutrition ou la mode.

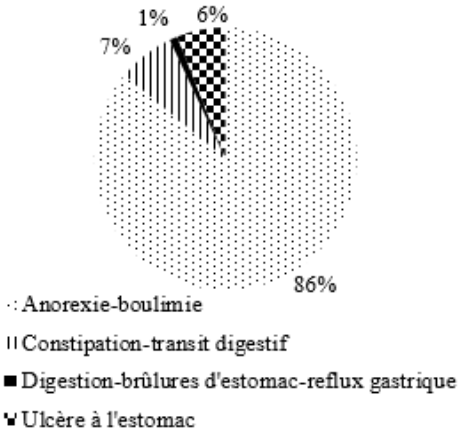


Figure 3. Constitution du corpus médical rédigé par des non-professionnels (Corpus Doctissimo)

Le corpus extrait de ce forum couvre 97 thèmes différents, 920.000 fils de discussion, plus de 23 millions de messages et 1,6 milliard de mots générés entre 2000 et 2017. Sur les 97 sujets initiaux, nous en avons sélectionné quatre (Figure 3) selon leurs liens avec la gastroentérologie, afin de créer une cohérence thématique avec le corpus iMediate. Le corpus utilisé pour cette étude est donc restreint à 76 000 fils de discussion, 1,8 million de messages et plus de 150 millions de mots.

3.1.2 Corpus de langue générale (Corpus CRF)

Le corpus de référence du français que nous comparons avec le corpus iMediate est un corpus que nous avons constitué nous-même à partir du principe de conception du Corpus de Référence du Français Contemporain (CRFC, Siepmann *et al.* 2016). Nous ne pouvions exploiter le CRFC en l'état actuel pour notre projet car il n'est pas encore rendu accessible ni téléchargeable et, conséquemment, parce qu'il n'est pas exploitable par nos différents logiciels – les requêtes doivent être réalisées en ligne, sur la

plateforme d'hébergement et de traitement de corpus SketchEngine (Kilgarriff *et al*. 2014).

C'est pourquoi nous avons constitué un corpus couvrant la même période (1950-2017) et reprenant les mêmes composantes du CRFC, mais de taille plus réduite et sans en respecter les proportions initiales.

Catégorie macro-générique	Sous-échantillon	Taille en millions de mots (CRFC)	Taille en millions de mots (CRF)
Oral	Interactions par oral spontané	30	1
Pseudo-oral	Pièces de théâtre et scénarios de film	30	1
	Sous-titres de films et de feuilletons télé quotidiens	2,5	1
	Textos	2,5	1
	Forums de discussion	60	1
Pseudo-écrit	Oral formel (allocutions, discours, informations)	30	1
Écrit	Universitaire et scientifique	30	1
	Romans et fiction en prose	30	1
	Journaux	45	1
	Journaux intimes et blogs	5	1
	Textes divers	4	1
Total		269	11

Figure 4. Composition du corpus CRF[4]

[4] Nous remercions particulièrement le centre de recherche VALIBEL (https://uclouvain.be/fr/instituts-recherche/ilc/valibel), le projet sms4science (http://www.sms4science.org/), le projet C-Phonogenre (Pršir, Goldman et Auchlin 2013) et le projet Orféo. (Debaisieux, Benzitoun et Deulofeu 2016) qui ont mis à notre disposition toutes leurs ressources linguistiques orales et pseudo-orales.

3.2 Méthodologie

Les corpus initiaux étant trop lourds pour être annotés par un humain ou pour être supportés par des outils soutenant et traitant l'annotation manuelle, nous avons constitué des échantillons de travail de chacun des trois corpus. Ces échantillons (Figure 5) incluent tout l'éventail de variabilité de la population, tout en respectant la proportion de chaque composant au sein de cette population (Biber 1993: 243-245).

Corpus iMediate	Corpus Doctissimo	Corpus de référence
13672	22040	62522

Figure 5. Taille des échantillons de travail (en nombre de mots)

En vue d'établir la typologie la plus précise possible, nous avons réalisé une analyse exploratoire de nos corpus consistant en l'annotation manuelle des sous-corpus précédemment constitués lors de deux sessions d'annotation réalisées par la même personne à trois mois d'intervalle[5].

	Corpus iMediate	Corpus Doctissimo	Corpus de référence
Annotation 1	818	1034	2452
Annotation 2	846	1067	2438
Total	955	1284	2259

Figure 6. Nombre d'annotations par corpus et par session d'annotation

Ces annotations ont été effectuées au moyen du logiciel d'annotation WebAnno (Yimam *et al.* 2014), adopté parce qu'il permet d'exporter les annotations sous la forme d'un format (.tsv) propre et exploitable par des outils informatiques et parce qu'il calcule automatiquement le coefficient kappa de Cohen, utile pour le calcul d'accord intra-annotateur.

[5] Nous ne pouvons donc rapporter de taux d'accord *inter*-annotateur, mais bien d'un taux d'accord *intra*-annotateur.

4. Établissement de la typologie

Afin d'établir cette typologie, nous avons procédé à un état de l'art des diverses typologies existantes en français. Nous présentons ci-dessous quelques-unes de ces références[6] et les modifications et ajouts que nous leur faisons sur la base de notre analyse exploratoire approfondie, ainsi que les éléments que nous conservons tels quels. Cette analyse exploratoire nous a ainsi permise d'enrichir à plusieurs niveaux les typologies préexistantes – notamment en ce qui concerne les spécificités médicales, de les raffiner et surtout d'exemplifier les dénominations théoriques au moyen d'occurrences extraites de nos corpus.

Les typologies causales de Jackiewicz (1998), Nazarenko (2000) et Gross (2009) constituent le fondement théorique de notre recherche. En effet, ces trois approches se combinent et offrent suffisamment d'informations comme point de départ scientifique pour dresser les prémices de notre propre typologie. Jackiewicz décrit précisément les constructions verbales causatives – peu abordées ailleurs ; Nazarenko propose et définit une typologie quadripartite qui intègre les connecteurs causaux, les connecteurs de conséquence et de but, les tournures syntaxiques à sens causal et le lexique ; Gross fournit une étude précise de la causalité sous toutes ses formes et nuances sémantiques en fournissant un essai de formalisation et en y incluant un panorama exhaustif de tous ses marqueurs. Toutefois, aucun n'apporte de moyen pour élucider les problèmes d'extraction liés à l'ambiguïté et à l'implicite.

Si des chercheurs ont remarqué l'existence d'une causalité implicite (ou non marquée, suivant les remarques de Blanco *et al.*[7]

[6] La notion de causalité a suscité énormément de travaux, tant en linguistique formelle, qu'en philosophie du langage ou en sémantique. Pour des raisons d'espace, il nous est impossible de citer tous les acteurs de cette effervescence, acteurs qui ont permis de nuancer, de préciser et de formaliser toutes les variétés d'expressions causales.

[7] Pour Blanco *et al.*, *marqué* (*explicite* pour nous, mais aussi ambigu) désigne les cas de causation « où il y a une unité linguistique spécifique qui signale la relation » ; *non marqué* (*implicite* pour nous) désigne les autres cas. Les concepts d'*explicite* et d'*implicite* renvoient quant à eux à la présence des

[2008: 310]), ils ne l'intègrent pas dans leur typologie ni n'offrent de moyens pour résoudre les difficultés d'identification ; il nous revient donc d'essayer de le faire. En outre, nous avons découvert qu'il existait davantage de types d'expressions causales ambigües (ce que l'on pourrait considérer comme des *implicatures* [Gazdar 1979]) que celles jusqu'ici enregistrées, comme les connecteurs logiques de concession, le *et* d'addition ou le lexique ambigu. Nous avons également incorporé à notre typologie (Figure 6) toutes les formes de causalité implicite que nous avons repérées dans nos corpus ainsi que les lexies médicales témoignant d'une relation causale.

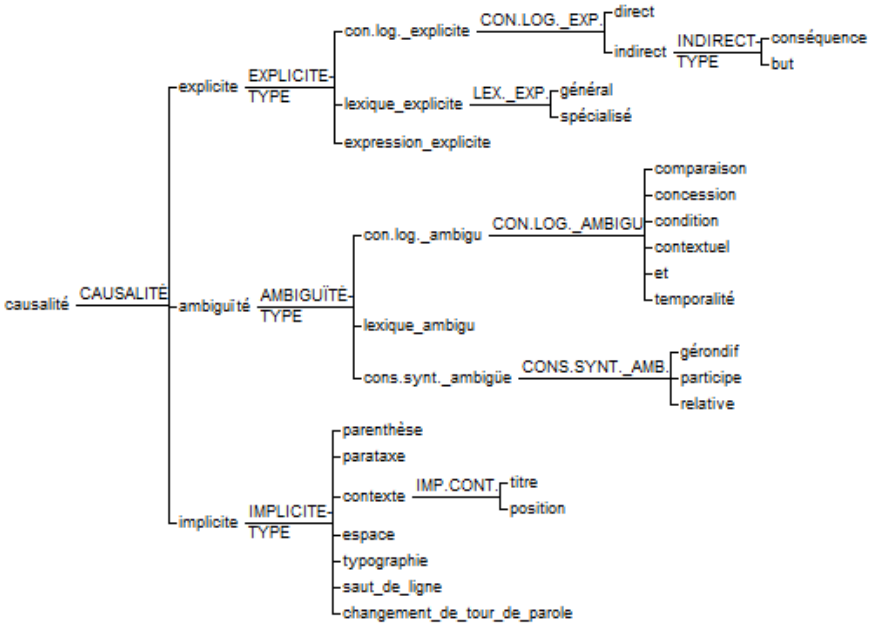


Figure 7. Typologie de la causalité établie pour les besoins du projet

deux arguments (*explicite*) que sont la cause et l'effet ou l'absence de l'un d'eux (*implicite*) dans l'expression de la relation causale.

Nous joignons à cette typologie une liste d'exemples pour aider à appréhender les différents types d'expressions causales que nous retenons dans notre modèle.

Explicite	Con.log._dir.	« Aucune CI au traitement nécessaire **car** pas d'angiome […] »
	Con.log._indir.csq.	« Le fond de la plaie […] a **donc** également été réséqué. »
	Con.log._indir.but	« […] on réalise également une acquisition tardive **afin de** bien visualiser les uretères. »
	Lex._gén.	« Un kyste sacro-coccygien […] **est responsable d'**un suintement persistant. »
	Lex._spéc.	« […] ulcère bulbaire hémorragique aigu, probablement **iatrogène**. »
	Expr.	« La morphologie du patient et la gravité de la péritonite **rendent** l'abord laparoscopique **impossible**. »
Ambiguïté	Con.log._comp.	« […] **plus** tu varies ton alimentation, **moins** t'as des problèmes de transit ! »
	Con.log._conc.	« La plaie, **bien que** large et profonde, est très propre […] »
	Con.log._cond.	« **Dans l'hypothèse d'**une éventuelle diarrhée sanglante d'origine infectieuse […] »
	Con.log._cont.	« Ces crises surviennent à chaque fois **dans** un contexte assez stressant. »
	Con.log._et	« […] je mangeais rien de la journée **et** le soir je me goinfri[ai] de cochonnerie en tout genre. »
	Con.log._temp.	« **Après** injection, on réalise également une acquisition tardive […] »
	Lex._ambig.	« […] élément qui **suggère** une irritation du péritoine. »
	Cons.synt._rel.	« [...] l'eau est retenue par mon corps **qui** a peur d'en manquer. »
	Cons.synt._part.	« […] **étant** fille unique, je comprends parfaitement ce que tu ressens. »
	Cons.synt._gér.	« […] c'est **en se restreignant** que ce sera pire. »
Implicite	Parataxe	« Je ne connais personne dans ma situation**,** je ne sais pas avec qui en parler […] »
	Cont._titre	« **Données cliniques :** Dolichocolon spastique connu »
	Cont._position	« Oesogastroduodénoscopie » (au tout début du document)
	Parenthèses	« 1 tasse de café avec du lait et deux sucre **(**j'arrive pas à boire le café sans sucre […]) »
	Espace	« […] je ne suis pas un modèle pour les repas je prends des 0% […] »
	Typographie	« […] je sens que j'ai quand même du chemin à faire **:** ça me semble ÉNORME ce que je mange […] »
	Saut_de_ligne	« L'anuscopie révèle un état inflammatoire de la muqueuse. Attitude proposée »

Figure 8. Exemples illustrant les différentes étiquettes possibles de la typologie

5. Observations

Cette mise à jour de typologies existantes pose deux problèmes qu'il nous faudra résoudre pour mener à bien la suite du projet. D'une part, en multipliant les types d'expressions causales, il devient parfois compliqué d'associer une seule étiquette à une expression (exemple 1).

(1) Antécédents médico-chirurgicaux appendicectomie vers 30 ans

Le segment « antécédents médico-chirurgicaux » peut s'interpréter soit comme un titre (« antécédents médico-chirurgicaux : appendicectomie ([…]) », marqué par 'implicite_contextuel_titre'[8]), soit comme une ambiguïté lexicale, avec l'idée qu'*antécédents* exprime potentiellement une cause, même si cette interprétation causale n'est pas la première signification de ce terme (ce segment serait estampillé par 'ambiguïté_lexique'[9]). En effet, en français, les *antécédents* sont des « affections antérieures à la maladie actuellement considérée » (TLF*i*, s.v. *antécédent*), sans aucune nuance de causalité. Afin de contourner ce problème de chevauchement entre catégories, il faut appliquer le principe de hiérarchie de la typologie, selon lequel l'explicite prime sur l'ambiguïté qui prime sur l'implicite, suivant la facilité de reconnaissance - et donc d'annotation – par un humain.

D'autre part, si nous, en tant qu'annotateurs humains, interprétons un texte et inférons des relations causales là où il n'y a pas de marque explicite de celles-ci, cela est presque impossible pour une machine ; or, nous avons pour objectif de développer un programme traitant correctement autant d'expressions causales du langage médical français que possible. Il y a donc encore des modifications et des améliorations à apporter à cette typologie, ainsi

[8] Cette étiquette indique un processus de causation qui n'est reconnue que par sa position d'en-tête aux différentes sections d'un fichier.
[9] L'ambiguïté causale concerne un mot notionnel, tel qu'un nom, un verbe ou un adjectif.

que des solutions à trouver – afin d'identifier les contextes propices à l'émergence d'une causalité implicite – avant d'élaborer un tel programme.

Nous implémentons actuellement une technique pour éviter le repérage systématique de causalité dans les cas d'ambiguïté et d'implicite, systématicité qui risquerait d'engendrer du « bruit »[10] dans nos extractions et, par conséquent, des erreurs dans nos analyses. Cette technique est constituée de deux étapes. D'abord, nous procédons à une analyse qualitative des contextes d'apparition des marqueurs de causalité ambigüe et implicite[11]. Pour ce faire, nous extrayons toutes les occurrences du marqueur considéré que nous annotons au minimum en causal/non-causal. Prenons par exemple le marqueur *quand* qui peut marquer la temporalité seule (*Ses parents ont divorcé quand il avait l'âge de 7 ans*, annotation [non-causal]) ou inclure une nuance causale (*Ses malaises surviennent quand elle a abusé de boissons alcoolisées* [causal]). Ensuite, nous appliquons la méthode de la classification automatique (Lenders 2006) qui permet, au départ de variables prédictives, d'indiquer si l'expression examinée ressort, en fonction de ces variables, plutôt de telle classe (causale) ou de telle autre (non-causale). Le défi de cette méthode repose sur l'identification méthodique de ces variables, dans notre cas, de variables syntaxico-sémantiques. Pour notre marqueur *quand*, voici deux exemples de variables prédictives et leur conséquence dans l'annotation (Figure 9).

[10] Par *bruit*, nous comprenons le *noise* anglais, à savoir des unités extraites mais pas pertinentes au niveau de l'application ou de l'objectif linguistique (L'Homme 2004: 192).

[11] La causalité implicite étant parfois indiquée par de la ponctuation ou d'autres marques typographiques, qui s'apparentent à des marqueurs dans le sens où on peut les repérer et les extraire automatiquement, par opposition aux espaces ou aux positions contextuelles.

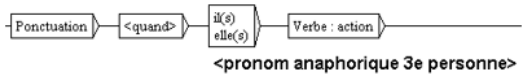
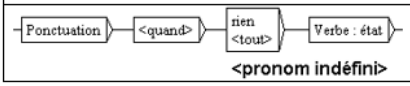
Variables : - syntaxique (position – pronom) - sémantique (type de procès)	Annotation	Exemple iMediate
	Causal	Ressent une douleur inguinale droite depuis quelques temps, **quand** elle porte son fils ou un poids […]
	Non-causal	En fait, la patiente m'a donné ton mot en fin de consultation, **quand** tout a été dit.

Figure 9. Exemples de variables prédictives de causalité ambigüe - quand

Il en va de même pour les cas d'implicite, du moins dans un premier temps. En effet, prenons l'implicite causal marqué par l'emploi de parenthèses. Il est évident que toutes les parenthèses introduites dans un rapport médical ne constituent pas des cas de causalité (entre autres fonctions, nous notons celles de « précision » ou d'« apport d'information secondaire non médicale »). Dans le cas des parenthèses (causalité implicite), l'expression d'une relation causale est plus subtile à appréhender– d'autant plus de manière automatique – ; il faut donc définir des variables syntaxico-sémantiques plus strictes encore, faisant appel notamment à la position de la parenthèse dans la phrase[12], à la catégorie grammaticale des termes qui l'entourent ou encore au type de verbe[13] de la principale dans laquelle la parenthèse est introduite.

[12] Il s'agit là d'une position relative calculée en fonction du nombre de mots séparant la fin de la parenthèse avec une ponctuation forte ; plus une parenthèse est proche de la fin de la phrase, plus on a tendance à l'interpréter comme causale.

[13] Pour établir les valeurs sémantiques des variables prédictives, nous nous référons aux tables du Lexique-Grammaire (Gross 1968) et aux Verbes français (Dubois et Dubois-Charlier 1994).

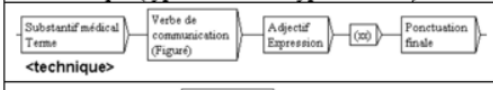
Variables : - syntaxique (position dans le texte) - sémantique (type de terme – type de verbe)	Annotation	Exemple iMediate
	Causal	La cytoponction s'est révélée rassurante (prolifération folliculaire).
	Non-causal	Par abord classique (médiane ou suspubienne) et coeloscopique.

Figure 10. Exemples de variables prédictives de causalité implicite - parenthèse

L'un des moyens les plus courants pour déterminer l'efficacité d'une typologie est de calculer l'accord intra-annotateur. En effet, calculer ce chiffre équivaut à se demander à quel point la typologie proposée pour la tâche d'annotation est suffisamment précise et objective pour laisser peu de place à la subjectivité, à l'interprétation et à la coïncidence chez un même locuteur à travers le temps.

Les deux dernières colonnes de la figure 11 exemplifient les divers cas que l'on peut rencontrer en confrontant les deux séries d'annotations : soit les annotations correspondent pour le segment annoté et l'étiquette associée ; soit les segments annotés sont identiques, mais les étiquettes divergent ; soit l'un des annotateurs n'a pas annoté l'expression causale relevée par l'autre annotateur.

Type	Segment annoté et position	Annotateur 1	Annotateur 2
Explicite	1302-1305 [tel]	Con.log._indir.csq.	Con.log._indir.csq.
Ambiguïté	12428-12443 [dans le sens où]	Con.log._cont.	Con.log._dir.
Implicite	1615-1616 [,]	Parataxe	Null

Figure 11. Exemples des trois cas pouvant survenir à la suite de l'inter-annotation

Suivant l'échelle de Santos (2015: 2), les coefficients kappa de Cohen offerts par WebAnno (figure 12) – mesurant l'accord intra-annotateur entre deux variables catégorielles – sont excellents. Ce haut accord peut s'expliquer par le fait qu'il s'agit du même annotateur lors les deux annotations et qu'il est encore largement en accord avec lui-même, même après un intervalle de trois mois. En outre, WebAnno ne tient compte que des annotations marquant les mêmes segments textuels, portant une étiquette identique (accord)

ou différente (désaccord), mais l'outil ne comprend pas les annotations qui diffèrent sur la taille du segment et sur la reconnaissance d'une expression causale.

	κ de WebAnno	κ recalculé par R
Corpus iMediate	0,96	0,71
Corpus Doctissimo	0,98	0,59
Corpus de Référence	0,97	0,58

Figure 12. Accords intra-annotateurs calculés par WebAnno – Accords intra-annotateurs recalculés par R

Afin de vérifier le calcul de l'accord, nous avons comparé les chiffres fournis par WebAnno avec ceux recalculés par R (R Development Core Team 2008 ; Figure 12) qui prend en considération les deux cas de figure rejetés par WebAnno (*cf.* note de bas de page 11). Cette légère différence d'accord intra-annotateur se comprend à la lumière de l'observation selon laquelle la causalité serait plus facile à détecter dans le langage professionnel qui repose davantage sur des normes et des règles de rédaction que la langue générale. Ceci est vrai même pour l'implicite, qui se situe souvent à des endroits déterminés et identifiables, tels que dans les titres de section des rapports (par exemple *Concerne* et *Indication*) ou dans les sauts de ligne.

6. Conclusion

De ces observations préliminaires, nous pouvons tirer les conclusions qui suivent :

(1) S'il est vrai qu'il nous reste du chemin à parcourir avant de pouvoir optimiser les systèmes d'extraction actuels, nous pensons qu'avoir délimité les différentes catégories d'ambiguïté et d'implicite est un progrès en soi qui nous permettra de développer les graphes nécessaires à l'élaboration d'un nouveau programme d'extraction, en ayant recours à des stratégies comme l'identification de

contextes propices à l'apparition de certains types d'expressions causales, l'utilisation d'un lemmatisateur ou d'un programme d'annotation des chaînes de coréférence.

(2) Pour évaluer une typologie ou un modèle, divers moyens sont disponibles. Au vu des résultats fournis par les accords intra-annotateur, nous remarquons qu'il est particulièrement complexe d'identifier (i) le segment textuel ou la part discursive qui communique une relation causale et (ii) le type d'expression causale dont il s'agit (quand elle n'est pas explicite). Notre typologie nous parait également à la fois suffisamment générale et granulaire pour notre recherche : elle prend en considération tous les types de discours (y compris l'oral, si nécessaire) et intègre des spécifications médicales (lexique spécialisé, titre et position, ce qui fait appel à des structures textuelles spécifiques au langage médical [Lehrberger 1982: 192-193]).

(3) Il nous faut établir un guide d'annotation précis, avec des règles d'application strictes afin d'augmenter l'efficacité de ce système et de résoudre ces deux problèmes. Il parait également opportun d'organiser une nouvelle campagne d'annotation avec deux annotateurs extérieurs à l'établissement de la typologie, afin de voir si les résultats obtenus en termes statistiques sont relativement similaires et s'il devient possible de développer des modèles informatiques d'extraction pour ensuite analyser linguistiquement ces données.

Références

Biber, D. (1993) ‘Representativeness in Corpus Design’. *Literary and Linguistic Computing* 8 (4), 243-257.

Blanco, E., Castell, N. et Moldovan, D. I. (2008) ‘Causal Relation Extraction’. *Proceedings of the sixth Language Ressources Evaluation Conference, LREC 2008 (26 mai - 1 juin 2008, Marrakech, Maroc)*, 310-313.

Bross, I.D.J., Shapiro, P.A. et Anderson, B.B. (1972) ‘How information is carried in scientific sub-languages’. *Science* 176, 1303-1307.

Corminboeuf, G. (2010) ‘La causalité sans les connecteurs ‘causaux’. Préalables épistémologiques’. *Linx* 62-63, 39-62.

Cougnon, L.-A. (2012) *L'écrit sms: variations lexicale et syntaxique en francophonie.* Thèse de doctorat, KU Leuven.

Debaisieux, J.-M., Benzitoun, C. et Deulofeu, J. (2016) ‘Le projet ORFÉO : un corpus d’étude pour le français contemporain’. In M. Avanzi, M. J. Béguelin et F. Diémoz, éds., *Corpus de français parlé et français parlé des corpus, Corpus* 15, 91-114.

Dubois, J. et Dubois-Charlier, F. (1994). *Les Verbes français*. Paris, Larousse.

Garcia, D. (1997) ‘COATIS, an NLP system to locate expressions of actions connected by causality links’. *Knowledge Acquisition, Modeling and Mangement, Proceedings of the Tenth European Workshop*, *EKAW ’97*, 347-352.

Gazdar, G. (1979) *Pragmatics. Implicature, Presupposition, and Logical Form*. New York, Academic Press.

Girju, R. et Moldovan, D. (2002) ‘Text Mining for Causal Relations’. *Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference*, 360-364.

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

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

Gross, M. (1968) *Grammaire transformationnelle du français. Vol. 1. Syntaxe du verbe*. Parsi, Larousse.

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

Lenders, W. (2006) ‘The Surface of Argumentation and the Role of Subordinating Conjunctions’. In A. Mehler et R. Köhler, éds., *Aspects of Automatic Text Analysis.* New York, Springer, 323-337.

Hirschman, L. et Sager, N. (1982) ‘Automatic Information Formatting of a Medical Sublanguage’. In R. Kittredge et J. Lehrberger, éds., *Sublanguage. Study of Language in Restricted Semantic Domains*. Berlin, New York, Walter de Gruyter, 27-80.

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

Itto, A. et Bouma, G. (2011) 'Extracting Explicit and Implicit Causal Relations from Sparse, Domain-Specific Texts'. In R. Munoz, A. Montoyo et E. Metais, éds., *Proceedings of the 16th International Conference on Applications of Natural Language to Information Systems*, *NLDB 2011*. 52-63.

Jackiewicz, A. (1998) *L'expression de la causalité dans les textes. Contribution au filtrage sémantique par une méthode informatique d'exploration contextuelle.* Thèse de doctorat, Université Paris-Sorbonne.

Khoo, C. S. G., Chan, S. et Niu, Y. (2000) 'Extracting Causal Knowledge from a Medical Database Using Graphical Patterns'. *ACL '00 Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 336-343.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovvář, V., Michelfeit, J., Rychlý, P. et Suchomel, V. (2014) 'The Sketch Engine: ten years on'. *Lexicography* 1, 7-36.

Kleinberg, S. et Hripcsak, G. (2011) 'A review of causal inference for biomedical informatics'. *Journal of Biomedical Informatics* 44 (6), 1102-1112.

Koch, P. et Oesterreicher, W. (2010) 'Gesprochene Sprache und geschriebene Sprache/Langage parlé et langage écrit'. In G. Holtus, M. Metzeltin et C. Schmitt, éds., *Lexikon der Romanistischen Linguistik* 1/2 : *Methodologie (Sprache in der Gesellschaft/Sprache und Klassifikation/Datensammlung und - verarbeitung)*. Tübingen, Niemeyer, 584-627.

L'Homme, M.C. (2004). *La terminologie. Principes et techniques.* Montréal, Presses de l'Université de Montréal.

Lehrberger, R. (1982) 'Automatic Translation and the Concept of Sublanguage'. In R. Kittredge et J. Lehrberger, éds., *Sublanguage. Study of Language in Restricted Semantic Domains.* Berlin, New York, Walter de Gruyter, 81-106.

Morlane-Hondère, F., Grouin, C., Moriceau, V. et Zweigenbaum, P. (2015) 'Médicaments qui soignent, médicaments qui rendent malades : étude des relations causales pour identifier les effets secondaires'. *TALN-RECITAL 2015 22ème conférence sur le Traitement Automatique des Langues Naturelles*, 22-25 juin 2015, Caen (France), 586-592.

Nazarenko, A. (2000) *La cause et son expression en français*. Paris, Ophrys.

Prsir, T., Goldman, J.-P. et Auchlin, A. (2013) 'Variation prosodique situationnelle : étude sur corpus de huit phonogenres en français'. *Prosody-Discourse Interface Conference 2013 (IDP-2013)*, Leuven.

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.

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. et Diwersy, S. (2016) 'Le Corpus de référence du français contemporain (CRFC), un corpus massif du français largement diversifié par genres'. *CMLF 5e Congrès Mondial de linguistique française* 27.

Sorgente, A., Vettigli, G. et Mele, F. (2013) 'Automatic extraction of cause-effect relations in natural language text'. *DART AI IA*, 37-48.

Sperlinga Gerner, M.-M. (2015) *Variations graphiques des textes des forums sur Internet*. Thèse de doctorat, Université de Strasbourg.

Watrin, P. (2006) *Une approche hybride de l'extraction d'information: sous-langages et lexique-grammaire.* Thèse de doctorat, KU Leuven.

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