that we have had access to in order to constitute our reference corpus

« UNE ZONE INFLAMMATOIRE *SÉQUELLAIRE* D'OESOPHAGITE PEPTIQUE »: FUNCTIONAL ANALYSIS OF THE EXPRESSION OF CAUSE/EFFECT PATTERNS IN MEDICAL RECORDS

Agathe PIERSON[1] Louise-Amélie COUGNON[2]

(1) PhD Student (FSR), CENTAL, IL&C, UCLouvain (2) FNRS Post-doc, CENTAL, IL&C, UCLouvain

INTRODUCTION

- The medical language is admitted as a *sublanguage* :
  "Certain proper subsets of the sentences of a language may be closed under some or all of the operations defined in the language, and thus constitute a sublanguage of it" (Harris 1971).

Sublanguage:
- Circumscribed subject area
- Common vocabulary
- Common habits of word usage

(Hirschman & Sager 1982)

→ Restrictions at the **syntactic**, **lexical** and **semantic** levels

- *Causality* is a phenomenon that reachs all these linguistic levels (Nazarenko 2002).
- *Causality* is everywhere in language because it helps to (Gross 2009):
  - dispel uncertainties
  - understand links between various events or states
  - structure the universe around us.

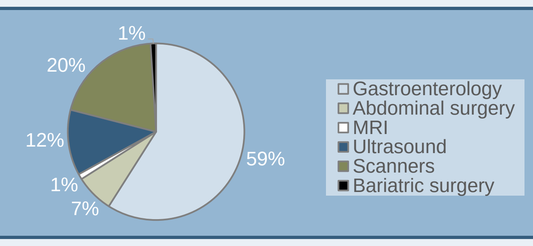
→ To consider causal expressions in medical language enables to enhance patient's safety, to take medical decisions and to improve the quality of provided care.

RESEARCH QUESTIONS

1. Are there **special features** relating to the use of causal expressions in medical language?
2. Are there **contextual factors** that have an influence on the use of a particular type of causal expression in written medical language?
3. Which **function** does each type of causal expression correspond to, according to these factors, in medical language?

CORPUS: iMediate corpus

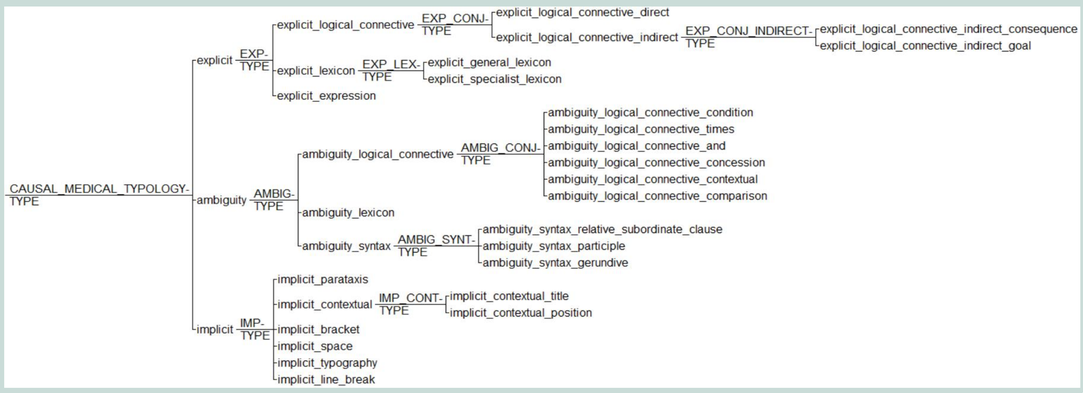
- 226,106 anonymised medical texts divided into 1,000 patient's medical files between 1996 and 2014
  → 88 million words
- Collected in six services of a Brussels hospital (gastroenterology, abdominal surgery, bariatric surgery, MRI, ultrasound and scanners)
- Gathering different sub-genres of medical texts (Patient Discharge Summary, Radiology Reports, etc.)



METHODOLOGY – SIX-STEP APPROACH

STEP 1

Exploratory analysis to set up our causal typology



STEP 2

Creation of a French reference corpus

on the basis of the composition of the CRFC (Siepmann, Bürgel & Diwersy 2016)

Informal spoken	Spoken
Stage plays and film scripts	Pseudo-spoken
Film and daily soap subtitles	
Text messages	
Discussion forums (Doctissimo)	
Formal oral (speeches, news)	Pseudo-written
Academic texts	Written
Non-academic texts	
Prose fiction	
Diaries and blogs	
Newspapers	

1 million words for each componant

↓

Sum: 11 million words

STEP 3

Comparison between medical professional language and general language

Manual annotation of one sample of both corpora (WebAnno 2014)

	Number of words	WebAnno	Edited by R
iMediate corpus	2,773	0.96	0.69
Complete reference corpus (CRC)	62,522	0.97	0.7
Reference corpus for (pseudo-)written language (RCWL)	25,413	0.98	0.74

Intra-annotator agreement

	General categories of causal expressions	Precise categories of causal expressions
iMediate corpus / CRC	χ^2 = 67.880 (p = 0.001) Cramer's V = **0.141** (p = 0.001)	χ^2 = 934.190 (p = 0.001) Cramer's V = **0.523** (p = 0.001)

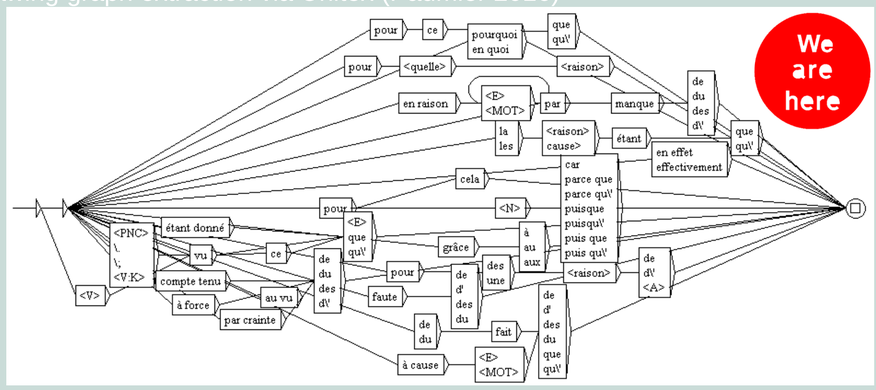
Chi-squared and Cramer's V between iMediate corpus and complete French reference corpus

	General categories of causal expressions	Precise categories of causal expressions
iMediate corpus / RCWL	χ^2 = 84.869 (p = 0.001) Cramer's V = **0.257** (p = 0.001)	χ^2 = 259.958 (p = 0.001) Cramer's V = **0.450** (p = 0.001)

Chi-squared and Cramer's V between iMediate corpus and French written reference corpus

STEP 4

Drawing graph extraction via Unitex (Paumier 2016)



Example of graph extraction of direct causal marker

STEP 5

Determine whether the 12 cpntextual factors that we consider have an impact on the types of causal expressions produced

Factor related to the medical staff
- *Prescriber's gender*

Factors related to the patient
- *Number of documents*
- *Age's patient*

Factors related to the disease
- *Type of disease*
- *Seriousness of the disease*
- *Temporality of the disease*

Factors related to the document
- *Sub-genre of the text*
- *Date of writing*
- *Place of writing*
- *Service*
- *Length of the text*

→ Test statistic between these contextual variables and the variable 'causal expression extracted'

STEP 6

Identification of the relevant factors and attemp to identify hypotetical functions for each type of causal expression depending on the context

	Categories of Explicit causal expressions
Services	χ^2 = 2473.022 (p = 0.001) Cramer's V = 0.169 (p = 0.001)

Test statistic for the variables Services and Explicit causal expressions

→ Conclusion: The variable « Services » is relevant but only 3% of the variance in type of explicit causal expressions may be explained as a result of the services from which come the files.

How to determine hypothetical functions?
- Look deeper at the results of the extractions of all causal expressions
- Try to notice anomalous uses from which we could venture hypothesis
- xxxxxxxxxxxx

CONCLUSION & PROSPECTS

Question research 1: Over: There are special features relating to the use of causal expressions in medical language → the variable "Type of corpus" explains 20% of the variability of the use of causal expressions (at the micro level)
→ This supports the requirement of a linguistic study of causal expressions in French medical language.

Question research 2: In progress: Require the awareness of methods of disambiguation and incorporation of implicit causal (October 2017)

Question research 3: In progress: Require graphs that extract all types of causal expressions (November 2017)
→ This will allow us to identify global functions (and no partial ones) in order to have a global insight of the functioning of the causal expression in French medical language

→ So as to confirm our hypothesis: field survey with a questionnaire (LimeSurvey) submitted to 150 physicians (Janvier 2018)

ACKNOWLEDGEMENT

Agathe Pierson is a Research Fellow of the Fonds Spécial de Recherche – FSR (Belgium)

We would also like to express our special thanks of gratitude to the research centre VALIBEL (UCLouvain), for its linguistic recourses to which we had access to constitute our reference corpus.

CONTACT

agathe.pierson@uclouvain.be
Faculté de philosophie, arts et lettres FIAL
Centre de Traitement Automatique du Langage (CENTAL)
Place Blaise Pascal 1/L3.03.12 B-1348 Louvain-la-Neuve (Belgium)
Tel. 32(0)10 47 37 77

15th International Pragmatics Conference, Belfast (Nothern Ireland), 16-21 July 2017