Contrastive study of syntactic phenomena in French written general and medical language corpora

Medical texts are precise and clear, but contain rich and complex terminology, as well as complex grammatical constructions. It is often necessary to simplify these texts to meet the needs of lay people. Hence, the simplification helps the patients to understand medical information and follow their treatment plans. Existing simplification guides provide general and imprecise indices on lexical, syntactic or semantic phenomena to be addressed. Besides, the linguistic phenomena that need to be simplified are not consensual in the guidelines.

Often, there is no empirical evidence of the real impact of these phenomena on understanding. Therefore, in order to propose the most suitable simplification, we must first study the real impact of such phenomena. In our study, we address the use of several syntactic constructions in medical and general-language texts in French. In this way, we want to examine whether these constructions are common in the general language corpora and, if so, their occurrence in medical texts may not impede the understanding.

We select four linguistic phenomena (passive voice, gerund, present participle, negation) and define their manifestations within corpora. We provide a contrastive study of these constructions in two French written corpora. Medical corpus contains articles from Wikipedia, and Cochrane documents available in the CLEAR corpus. We randomly sample 404 texts. For the general-language corpus, we chose 446 articles, written in 2013 in the journal Le Monde. These articles cover different topics (politics, culture, fashion, science, etc.).

In order to collect evidence on the real use of the syntactic constructions, we carry out a quantitative approach and conduct an observational comparative study. Hence, we apply NLP tools to determine the parts of speech (POS) and grammatical characteristics of words, as well as dependency and morphological analysis, using the spaCy library in Python. Then, the extraction rules are applied to detect sentences with the target phenomena. For example, the rule for identifying present participle is *[{'POS': 'VERB', 'MORPH': 'Tense=Pres|VerbForm=Part}]*, where sequences POS-tagged as verbs and having morphological features *present time* and participle are searched for. The example for this construction is for instance: *certains patients souffrant de maladies inflammatoires chroniques (some patients suffering from chronic inflammatory diseases).*

We have automatically processed 52,364 sentences (623,155 occ.) from medical texts and 19,923 sentences (293,427 occ.) from general-language texts. As part of the results, we indicate here what we obtain on present participle and passive voice. (1) For present participle, the mean average frequency per sentence is 10,6% in medical texts and 8,4% in general-language texts. The standard deviation in general-language texts is 0.072 (0 to half of the sentences per text). In medical texts, the standard deviation is 0.088 (0 to 63% of the sentences per text). (2) For passive voice, the mean average frequency per sentence is 18.3% in medical texts and 8.01% in general-language texts. The standard deviation in medical texts is 0.147 (0 to 77% of sentences per text), while the standard deviation in general-language texts is 0.065 (0 to 43% of the sentences per text). These results are statistically significant, with the estimated p-value $< 10^{-3}$. Even though the frequency of passive voice is twice as high in medical texts by comparison with general-language texts, we observe that these constructions occur quite frequently in general texts as well.

A more complete analysis will be provided, consisting of additional insights in these syntactic phenomena, such as: breakdown of syntactic phenomena per construction, which passive constructions are used more often, in which grammatical tense, difference according to text genre, etc. Other linguistic phenomena addressed will also be presented.

Bibliography

Altinok, D. (2021). *Mastering spaCy: An end-to-end practical guide to implementing NLP applications using the Python ecosystem*. Packt Publishing Ltd.

Colic, Nico, and Fabio Rinaldi. "Improving spaCy dependency annotation and PoS tagging web service using independent NER services." *Genomics & informatics* 17.2 (2019).

Grabar N, Cardon R. CLEAR – Simple Corpus for Medical French. In : Workshop on Automatic Text Adaption (ATA) ; 2018. p. 1–11.

Javourey-Drevet, L., Dufau, S., François, T., Gala, N., Ginestié, J., & Ziegler, J. (2022). Simplification of literary and scientific texts to improve reading fluency and comprehension in beginning readers of French. *Applied Psycholinguistics, 43*(2), 485-512. doi:10.1017/S014271642100062X

Laetitia Brouwers, Delphine Bernhard, Anne-Laure Ligozat, Thomas François. Simplification syntaxique de phrases pour le français. Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, Jun 2012, Grenoble, France. pp.211-224. ⟨hal-00790862⟩

Miller, T., Leroy, G., Chatterjee, S., Fan, J. & Thoms, B. (2007). A classifier to evaluate language specificity of medical documents. In HICSS, pp. 134–140.

Tchami, Ornella Wandji, N. Grabar and Ulrich Heid. “Syntagmatic Behaviors of Verbs in Medical Texts: Expert Communication vs. Forums of Patients.” *International Conference on Terminology and Artificial Intelligence* (2015).

UNAPEI. L’information pour tous. UNAPEI ; 2019. Available from : https://easy-to-read.eu/wpcontent/uploads/2014/12/FR_Information_for_all.pdf.

Eleni Partalidou, Eleftherios Spyromitros-Xioufis, Stavros Doropoulos, Stavros Vologiannidis, and Konstantinos Diamantaras. 2019. *Design and implementation of an open source Greek POS Tagger and Entity Recognizer using spaCy.* In IEEE/WIC/ACM International Conference on Web Intelligence (WI '19). Association for Computing Machinery, New York, NY, USA, 337–341. https://doi.org/10.1145/3350546.3352543