Analysis and Automatic Processing of Discourse

Manon Cassier (*AGORA*), Julien Longhi (*AGORA*), Damien Nouvel (*ERTIM*), Agata Jackiewicz (*PRAXILING*), Jean-Yves Antoine (*LIFAT*), Anaïs Lefeuvre-Halftermeyer (*LIFO*).

The ANR TALAD project aims to demonstrate the added value of Discourse Analysis (DA) and Natural Language Processing (NLP) for collaborative work. It focuses on "nominations" (Moirand, 2011; Frath, 2015; Longhi, 2015) - a concept commonly investigated in DA. This concept refers to the way in which speakers name things with a meaning that contains some semantic appreciation (like an opinion) - typically by reusing denominations (Kleiber, 1984) that already exist (for instance *migrant* versus *refugee* (Alsadhan *et al.*, 2018)) but whose meaning is somehow updated by their context of appearance. The new meaning entirely depends on the speaker's deictics and provides access to his ideology (Veniard, 2013; Lecolle, 2016; Jackiewicz & Pengam, 2018).

The project aims, on the one hand, at examining existing literature in DA to help characterize the concept and, on the other, to show what NLP contributions could be proposed to easily detect and analyze nominations. The first work will lead to the construction of an online ontology of theoretical concepts from both fields (DA and NLP) and references that will be useful for the study of nominations. This ontology will allow us to build an annotation guide that will be used to annotate a French corpus of political interviews' transcripts (Chilton, 2004; Van Dijk, 1997, 2002) that contain nominations. It will serve as a reference for the DA community (Baker, 2006). It will also be used to test or implement approaches that automate nomination extraction and analysis. The project aims at defining interpretative paths based on schemas related to the semantics of the entity referred to by the nomination.

The corpus (collected by a private company) for now contains manual transcripts of political interviews given during multiple morning radio broadcasts from October 2016 to December 2017 - a period which encompasses the 2017 Presidential French election. It contains more than 3,000 transcripts with 10 million tokens, related to more than 500 interviewed public figures. For each interview, metadata specifies the names of the interviewer and interviewee. Keywords have been manually provided along with transcripts related to topics covered by the interviews. However, they greatly vary depending on the annotator: their quality and subjectivity are questionable. We plan to use this annotated corpus to observe nominations in context and establish rules or features to automatically detect them.

The available definitions for the nomination process in existing literature lead our study towards the use of entity recognition (Ehrmann *et al.*, 2016; Nouvel *et al.*, 2012) and coreference chain analysis (Désoyer *et al.*, 2016) systems. The recognition and disambiguation of entities, by linking linguistic content to what they refer to, should provide possible variations for a single entity. Similarly, the study of coreference chains (as identified by an automatic coreference resolution system) would give useful insights into their lexical variants. More generally, we expect both to make it possible to better understand the intentional diversion of meaning they involve on denominations.

We believe that this project would benefit DA and NLP with respect to theoretical, methodological and practical aspects. In particular, we hope that the conclusions may help the two communities to distinguish referential (designation) and axiological (opinion) components of language.

References:

Alsadhan, M., Jackiewicz, A., & Luxardo, G. (2018, June). Migrants et réfugiés: dynamique de la nomination de l'étranger. In *JADT'18: Proceedings of the 14th International Conference on Statistical Analysis of Textual Data*.

Baker, P. (2006). *Using corpora in discourse analysis*. A&C Black.

Chilton, P. (2004). Analysing political discourse: Theory and practice. Routledge.

Désoyer, A., Landragin, F., Tellier, I., Lefeuvre, A., Antoine, J. Y., & Dinarelli, M. (2016, April). Coreference Resolution for French Oral Data: Machine Learning Experiments with ANCOR. In *International Conference on Intelligent Text Processing and Computational Linguistics,* 507-519. Springer, Cham.

Ehrmann, M., Nouvel, D., & Rosset, S. (2016). Named entity resources-overview and outlook. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, 3349-3356.

Frath, P. (2015). Dénomination référentielle, désignation, nomination. *Langue française*, (4), 33-46.

Kleiber, G. (1984). Dénomination et relations dénominatives. *Langages*, (76), 77-94.

Jackiewicz, A., & Pengam, M. (2018, June). Des «musulmans modérés» dans les discours médiatiques. Étude linguistique d'une expression controversée. In *Les représentations médiatiques de l'islam et des musulman.es.*

Lecolle, M. (2016). Noms collectifs humains : nomination et prédication. *Argumentation et Analyse du Discours*, (17).

Longhi, J. (2015). Stabilité et instabilité dans la production du sens : la nomination en discours. *Langue française*, (4), 5-14.

Moirand, S. (2011). Du sens tel qu'il s'inscrit dans l'acte de nommer. *Humanitas / Fapesp Ciências da linguagem e didatica das linguas,* p.165-180.

Nouvel, D., Antoine, J. Y., Friburger, N., & Soulet, A. (2012, April). Coupling knowledge-based and data-driven systems for named entity recognition. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, 69-77. Association for Computational Linguistics.

Van Dijk, T. A. (2002). Political discourse and ideology. *Anàlisi del discurs polític*, 15-34.

Van Dijk, T. A. (1997). What is political discourse analysis. *Belgian journal of linguistics*, 11(1), 11-52.

Veniard, M. (2013). *La nomination des événements dans la presse. Essai de sémantique discursive.* Presses universitaires de Franche-Comté.