# Analysis and Automatic Processing of Discourse

Manon Cassier, Julien Longhi, Damien Nouvel, Agata Jackiewicz, Jean-Yves Antoine, Anaïs Lefeuvre-Halftermeyer

# Analysis and Automatic Processing of Discourse

Manon Cassier[1 2], Julien Longhi[1], Damien Nouvel[2], Agata Jackiewicz[3], Jean-Yves Antoine[4] and Anaïs Lefeuvre-Halftermeyer[5]

[1] Université de Cergy-Pontoise, AGORA
[2] Institut National des Langues et Civilisations Orientales (INALCO), ERTIM
[3] Université Paul Valéry - Montpellier 3, Praxiling
[4] Laboratoire d'Informatique Fondamentale et Appliquée de Tours (LIFAT)
[5] Laboratoire d'Informatique Fondamentale d'Orléans (LIFO)

## Abstract

The **ANR TALAD** project aims to demonstrate the added value of Natural Language Proccesing (NLP) for Discourse Analysis (DA) works. It focuses on the **nomination** DA concept for :

- the construction of a reference thesaurus for the annotation of DA concepts,
- the annotation in nominations of a french political interviews corpus,
- a system modelling for automatic recognition of nominations.

## 1. Motivations

Some **textometric tools** (TXM [4], Iramuteq [8], Lexico3 [7]) are already used to assist French DA but :

- Do not allow an automatic extraction of DA concepts
- Only provide topic classification and do not display how these topics are discussed (i.e. how the speaker expresses his opinion)

## 2. The nomination concept

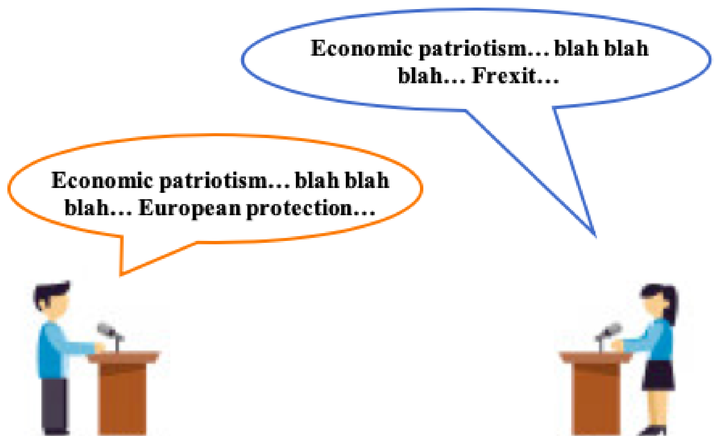
- Designates the **updating of an entity's identity** [3] :
  - → By the speaker, in discourse
  - → Through its use of the name (context or name itself)
- Usually concerns social entities and is generally stimulated by a conflicting political or social context [5][9]
- Gives an indication on the speaker's bias and ideology. [6]

**Ex.** "Si la France veut continuer de se projeter dans le monde, elle doit avancer en **Europe**, rebâtir le **projet européen** avec détermination et là aussi ne rien céder à celles et ceux qui doutent."

→ If France wants to continue to project itself into the world, it must move forward in Europe, rebuild the European project with determination and, here too, give nothing away to those who doubt.

**Ex.** "Donc l'**Union européenne** est une mauvaise chose, pas l'**Europe**. L'**Europe** c'est une entité, une réalité de civilisations, c'est une réalité géographique, historique."

→ So the European Union is a bad thing, not Europe. Europe is an entity, a civilization, a geographical and historical reality.

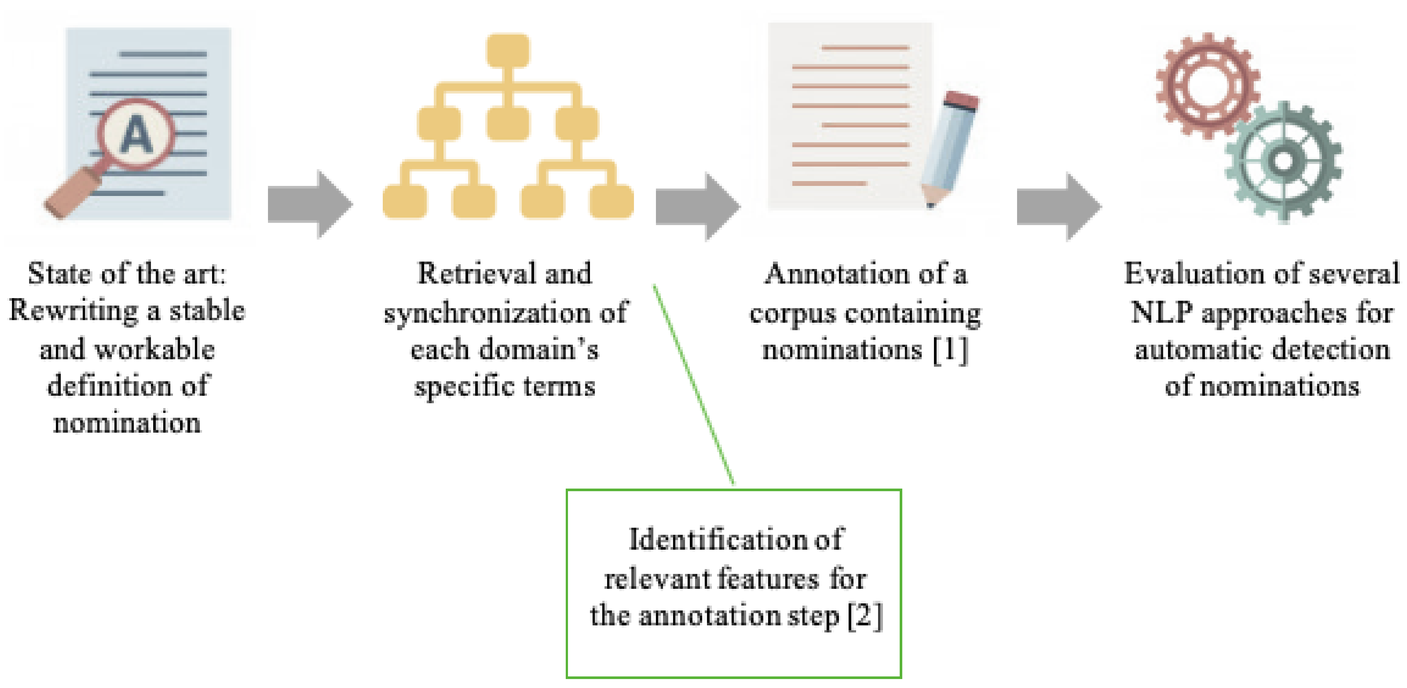



## 3. Corpus

As nominations are generally motivated by political conflicts (and then reused in common language) our corpus contains :

- **3168 manual transcripts of interviews** (more than 10M tokens) of politicians invited to French radio broadcasts
- Recorded during the 2017 presidential elections (October 2016 to December 2017)

## 4. Working steps




## 5. Further Research

- Recognizing and disambiguating entities - by linking linguistic content to what they refer to - should provide possible variations for a single entity.
- Coreference chains - as identified by an automatic coreference resolution system - would also give useful insights into their lexical variants.
- Word embeddings also seem interesting to model the semantic variation of a name through the corpus.

## Selected References

1 Baker, P. (2006). Using corpora in discourse analysis. A&C Black.
2 Chilton, P. (2004). Analysing political discourse : Theory and practice. Routledge.
3 Frath, P. (2015). Dénomination référentielle, désignation, nomination. Langue française, (4), 33-46.
4 Heiden, S. (2010). The TXM Platform : Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In K. I. Ryo Otoguro (Ed.), 24th Pacific Asia Conference on Language, Information and Computation (p. 389-398). Institute for Digital Enhancement of Cognitive Development, Waseda University.
5 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.
6 Van Dijk, T. A. (2002). Political discourse and ideology. Anàlisi del discurs polític, 15-34.
7 Lamalle, C., Martinez, W., Fleury, S., Salem, A., Fracchiolla, B., Kuncova, A., & Maisondieu, A. (2002). Lexico 3, Outils de statistique textuelle. Manuel d'utilisation. Université de la Sorbonne Nouvelle.
8 Ratinaud, P. (2009). *Iramuteq : Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*. http ://www.iramuteq.org
9 Veniard, M. (2013). La nomination des événements dans la presse. Essai de sémantique discursive. Presses universitaires de Franche-Comté.