4L7TL02P - Modélisation linguistique pour l'analyse automatique de textes

Introduction

Delphine Battistelli
delphine.battistelli@parisnanterre.fr

Université Paris Nanterre
22 septembre 2022

Organisation du cours

Espaces numériques

- Espace cours en ligne Nanterre : https://coursenligne.parisnanterre.fr/course/view.php?id=5296 (clé du cours : MAAT_20222023)

À installer sur vos machines

- Logiciel Unitex : https://unitexgramlab.org/fr
- Logiciel Glozz : http://www.glozz.org/

Organisation du cours

La modélisation en linguistique et en TAL – Plan de cours

- Introduction théorique
- Exemple applicatif 1 : les adverbiaux temporels
- Exemple applicatif 2 : les émotions

Organisation du cours

Évaluation

- DM Unitex et adverbiaux temporels – 40% (mi-novembre)
- Partiel final – 60% (dernière semaine avant vacances Noel)

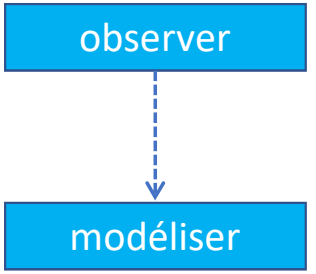
Introduction :
Modélisation linguistique

Qu'est-ce que "modéliser" ?

– La démarche scientifique consiste à observer, décrire, classer, généraliser, modéliser

  • L'*observation* est la toute première phase de la démarche scientifique,

  • tandis que la *modélisation* en constitue la phase ultime



– Ceci est vrai en linguistique comme dans d'autres sciences

  • Chomsky, dans les années 1950, a ainsi explicitement voulu donner à la linguistique le statut de science à l'égal des sciences de la nature

Qu'est-ce que "modéliser" ?

Démarche scientifique

Observation

↓

Description

↓

Classement

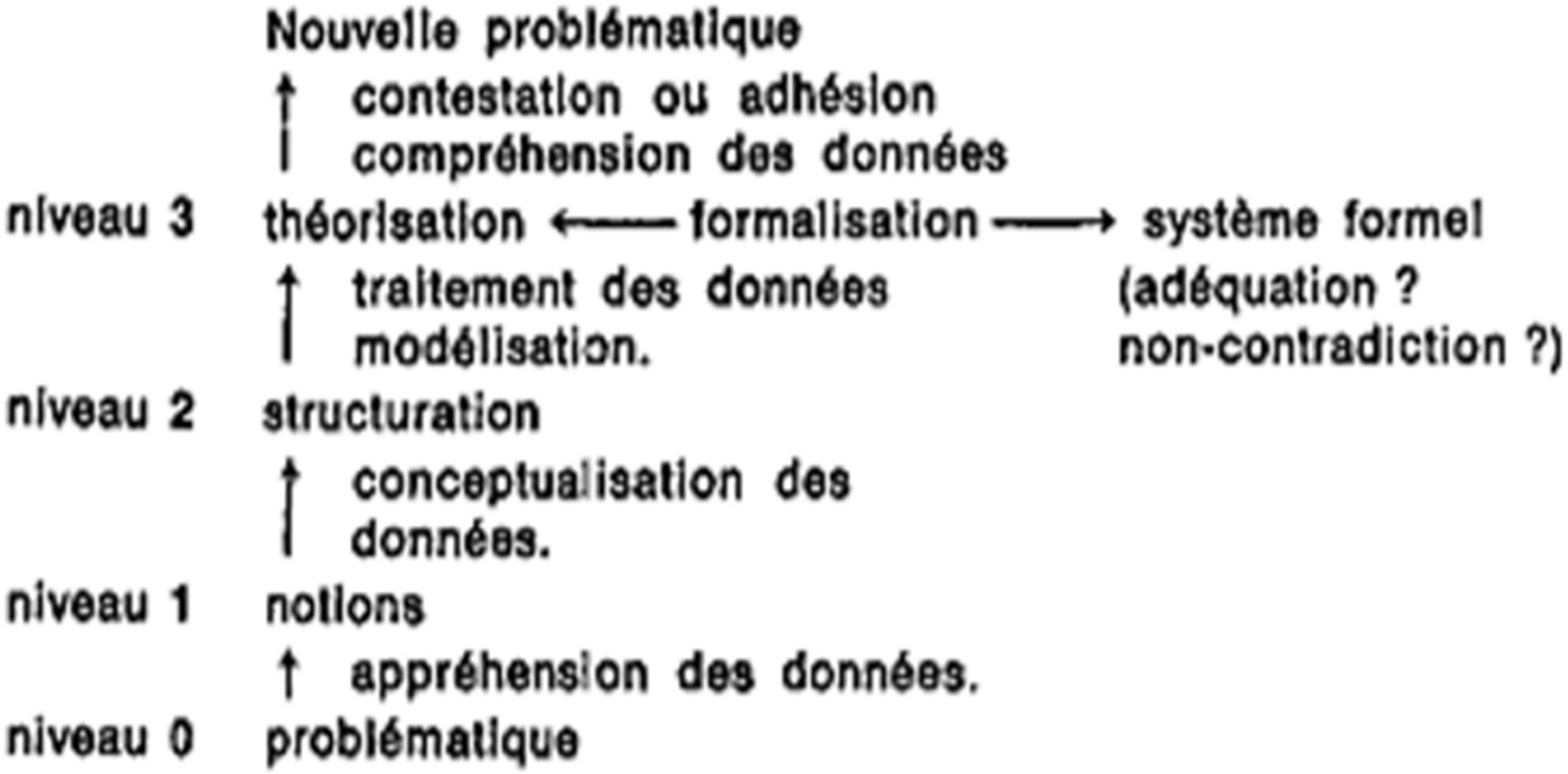
↓

Généralisation

↓

Modélisation

Modéliser en linguistique



In Desclés J.-P. (1969), Linguistique et Mathématiques, l'Homme, 9-3, pp. 93-99

Modéliser en linguistique

- Modéliser pour rendre compte de phénomènes observables
- Mesurer l'adéquation d'une théorie aux faits de langue qu'elle prétend expliquer
- Objectif : compromis entre simplicité du système *et* adéquation aux données observées
- Mise à contribution des mathématiques grâce aux outils informatiques dans la construction des modèles

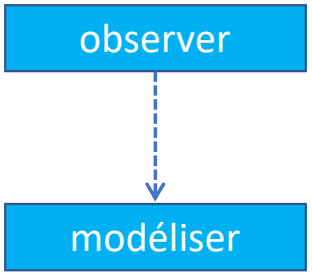
"Modéliser" en TAL ?

A LIRE :

HABERT, Benoît. « Portrait de linguiste(s) à l'instrument ». *Texto!* [en ligne], décembre 2005, vol. X, n°4
http://www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html

VICTORRI, Bernard. « Le modèle en linguistique », *Encyclopaedia Universalis*, 1997
Version préliminaire disponible sur
http://halshs.archives-ouvertes.fr/halshs-00009518

Les observables en TAL

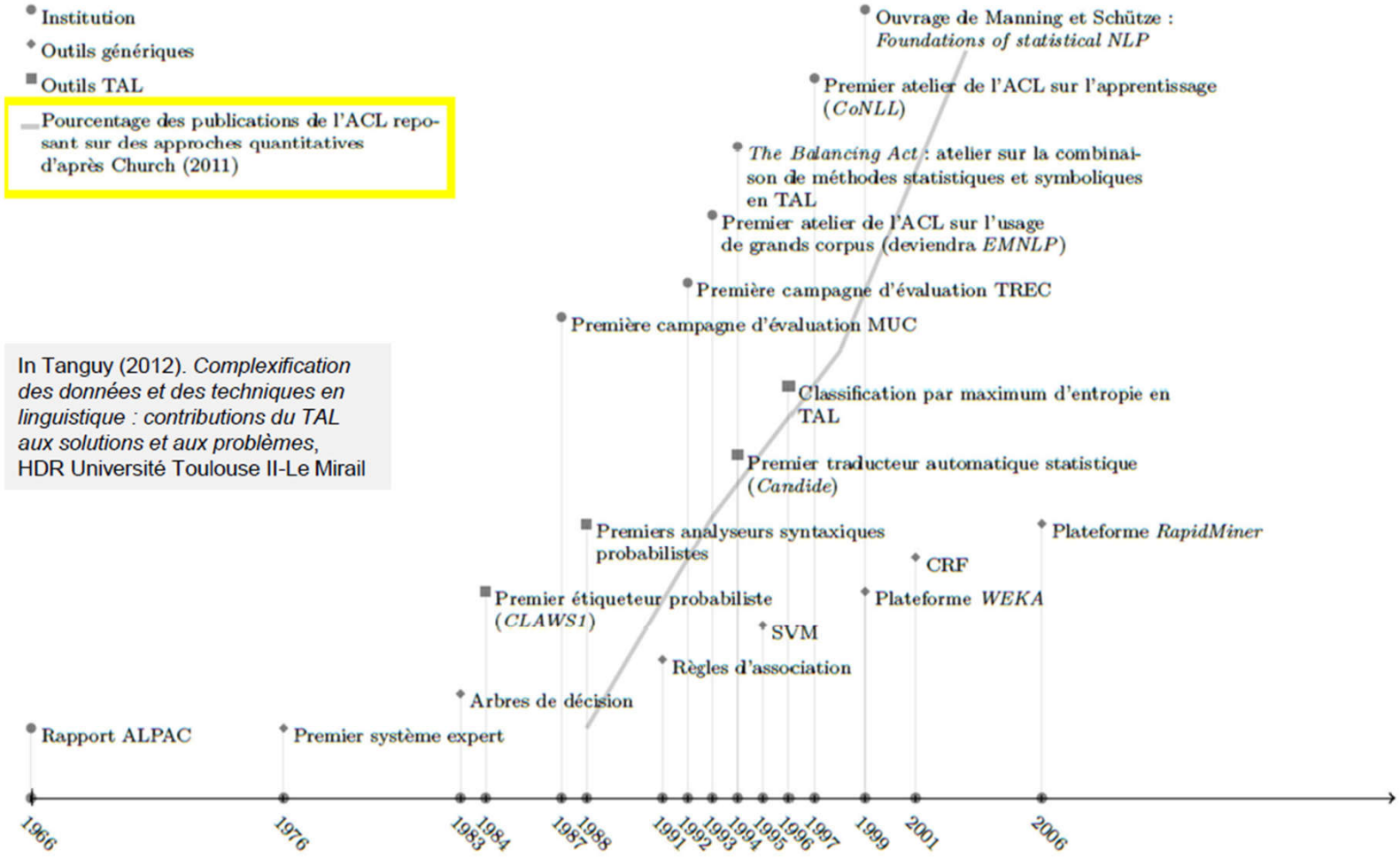


- TAL avant 2000' : recours très limité à des corpus et appel à l'intuition
- TAL depuis 2000' :
  - un **changement d'échelle** dans les **observables** : des corpus très volumineux, et aussi très diversifiés (articles scientifiques, blogs, romans, tweets, …)
  - Développement d'un nouveau paradigme : les **corpus annotés manuellement** et les **méthodes par apprentissage** (*i.e.* méthodes statistiques) : dans ces méthodes, « le rôle principal attribué au linguiste (…) se limite à l'annotation des données, que ce soit à des fins d'entraînement ou d'évaluation » (Fabre, 2010)

-> *Observe*-t-on vraiment les données ? Quelle est la place du linguiste ? -> Quelle est la place des outils de TAL dans cette observation ?

-> Le TAL a-t-il encore besoin des linguistes ?

"Modéliser" en TAL ?



In Tanguy (2012). *Complexification des données et des techniques en linguistique : contributions du TAL aux solutions et aux problèmes*, HDR Université Toulouse II-Le Mirail

Modélisation linguistique et TAL

Le TAL a-t-il encore besoin de linguistes ?

- Oui !
- Performances limitées aux phénomènes fréquents (Hajicova, 2011), pourtant phénomènes "rares" nombreux cf. loi de Zipf (Kay, 2011)

Hajičová, E. (2011). Computational linguistics without linguistics ? view from prague. Linguistic Issues in Language Technology, 6.
Kay, M. (2011). Zipf's law and l'arbitraire du signe. Linguistic Issues in Language Technology, 6.

Le TAL a-t-il encore besoin des linguistes ?

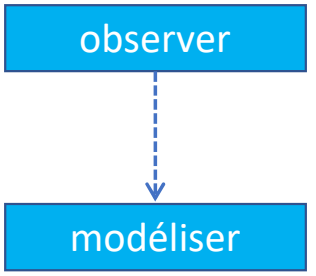
La réponse est oui !

- Kay (2011) et Hajicová (2011) rappellent que les connaissances linguistiques pointues sont les seules façons de faire progresser les systèmes au-delà de la zone des phénomènes suffisamment répétés, *i.e.* les fameux 80 premiers %, alors que la loi de Zipf nous a depuis longtemps appris qu'un grand nombre de phénomènes très rares constituent la partie manquante de la couverture de tout système de TAL

- Dans tous les cas, Church (2011) prédit un retour cyclique du balancier vers des travaux de TAL plus théoriques et moins empiriques lors de la prochaine décennie, avec dans ce cadre une remontée en puissance de la linguistique sur la scène du TAL

Church, K. (2011). A pendulum swung too far. *Linguistic Issues in Language Technology, 6.*
Hajičová, E. (2011). Computational linguistics without linguistics ? view from prague. *Linguistic Issues in Language Technology, 6.*
Kay, M. (2011). Zipf's law and l'arbitraire du signe. *Linguistic Issues in Language Technology, 6.*

Les principes adoptés dans ce cours



- Le TAL a besoin de **modèles linguistiques** rigoureux à implémenter pour optimiser les applications

=> On adopte une démarche de modélisation linguistique et l'on met à contribution des mathématiques grâce aux outils informatiques dans la construction d'un modèle

Exemple : la notion (intuitive) de « date »

- En linguistique : renvoie à la notion d' « adverbe temporel »
- En extraction d'information : est qualifiée d' « entité nommée »
- En TAL : dans les classifications (annotations) proposées, on opposera le plus souvent « date », « durée » et « fréquence » et on s'intéressera également à l'opposition relatif *vs* absolu

-> modéliser cette notion *linguistiquement*
- *cf.* TD1

Exemple applicatif 1 :
Les adverbiaux temporels