Ressources textuelles incrémentales pour la modélisation des interactions linguistiques multiples

Maria Zimina, Serge Fleury

Zimina, Maria & Serge Fleury
University Paris Diderot, France & Université de la Sorbonne Nouvelle – Paris 3
mzimina@eila.univ-paris-diderot.fr
serge.fleury@univ-paris3.fr

TITRE
Ressources textuelles incrémentales pour la modélisation des interactions linguistiques multiples

MOTS-CLÉS
analyse de corpus annotés, modélisation linguistique, ressources textuelles incrémentales

RÉSUMÉ
Le besoin d'interroger des données textuelles qui reflètent des phénomènes langagiers variés et structurellement hétérogènes existe dans plusieurs domaines de recherche en linguistique appliquée. Face à ces enjeux, l'utilisation d'outils de constitution et d'analyse de corpus textuels est devenue incontournable (Hyland *et al*., 2012). Ces outils reposent sur des hypothèses directement liées aux résultats d'expérimentations. Dans ce contexte, les questions de codage des données, les types d'unités de base utilisés dans les calculs statistiques, la pertinence des ressources et modèles linguistiques mobilisés revêtent une importance capitale.

Afin d'aboutir à la représentation coordonnée de multiples phénomènes observables à plusieurs niveaux d'analyse linguistique, l'approche la plus courante consiste à utiliser des environnements intégrés d'expérimentation et d'annotation automatique. L'étude des fonctionnements interactionnels est alors envisagée au sein des plateformes unifiées [1] qui s'orientent vers la standardisation de l'annotation des données textuelles et leur exploitation à l'aide de langages de requêtes. Cette approche a certaines limites. Elle rend difficile la systématisation des comparaisons des objets linguistiques et des liens qui existent entre eux à plusieurs niveaux d'analyse. De plus, les objets construits à l'aide de requêtes sont de nature variable et ne peuvent pas être confondus dans les mêmes décomptes au cours de l'analyse quantitative.

Pour préciser d'avantage les apports croisés entre qualitatif et quantitatif, nous menons une réflexion sur la modélisation systémique de la structure du matériau textuel informatisé. Cette réflexion s'appuie sur les résultats de développement d'un modèle de données concret qui s'inspire des avancées récentes de l'analyse de données textuelles. Pour analyser un corpus de textes électronique, on construit un système de décompte d'unités résultant du processus de segmentation automatique. Le flux textuel se présente alors sous forme d'une succession d'*items* numérotés qui fournissent un système de coordonnées sur le texte : la *Trame*. Les empans textuels (parties) sont indexés sur la *Trame* comme suites d'*items* consécutifs, entre la position x_1 et la position x_2. Les systèmes d'empans sont regroupés dans une structure de données appelée *Cadre*.

Une ressource textuelle constituée sous la forme *Trame*/*Cadre* est utilisée pour un repérage des objets type *Sélections* qui peuvent être soumis à l'analyse quantitative. Les *Sélections* de *contenus* sont des *items* correspondant aux occurrences d'un *type* (forme, lemme, patron morphosyntaxique,

[1] GATE (General Architecture for Text Engineering) : http://gate.ac.uk/gate
ANNIS (ANNotation of Information Structure) : http://annis-tools.org/
MACAON (chaîne de traitement) : http://macaon.lif.univ-mrs.fr/

expression régulière croisant plusieurs annotations). Les *Sélections* de *contenants* sont constitués d'*items* connexes (zones, parties, sections, paragraphes). Indexées sur une *Trame* commune, les *Sélections* sont analysées au sein des tableaux croisant les décomptes de chacun des *types* (*contenus*) dans chacune des parties (*contenants*). Elles sont transmises entre procédures de traitement.

Notons que ce modèle de données permet de stocker non seulement les découpages du texte mais aussi les annotations produites par les différentes procédures informatiques et, éventuellement, les passer d'une procédure de traitement à l'autre. Par conséquent, l'annotation s'intègre dans le processus dynamique d'exploration : elle est créée, prise en compte ou corrigée dans le C*adre* défini à partir d'une T*rame* unique (Fleury, Zimina, 2014).

En suivant ces principes, des comparaisons entre des ressources textuelles issues de cadres théoriques multiples mobilisent les procédures de navigation en corpus. Les états successifs de traitement induisent la notion de *ressource textuelle incrémentale* qui conserve la trace de séquences de traitement apportées à la ressource textuelle initiale. En suivant ces principes, il devient possible d'observer les apports respectifs de plusieurs formalismes, se pencher sur les modes de leur utilisation conjointe en analyse de corpus, leurs objectifs respectifs et les différences dans les résultats d'expérimentions qu'elles amènent.

Nous illustrons cette approche à l'aide des données du corpus *Rhapsodie*[2] (corpus prosodique de référence en français parlé). La démonstration mettra en évidence les spécificités des genres textuels en s'appuyant successivement sur des unités décrites par des annotations textuelles multiples (prosodiques, syntaxiques, etc.).

RÉFÉRENCES

Serge Fleury, Maria Zimina (2014). Trameur: A Framework for Annotated Text Corpora Exploration. *COLING 2014, the 25th International Conference on Computational Linguistics. System Demonstrations*, August 2014, Dublin, Ireland.

Serge Fleury (2013). *Le Trameur. Propositions de description et d'implémentation des objets textométriques*. Sorbonne nouvelle – Paris 3. En ligne : <http://www.tal.univ-paris3.fr/trameur/trameur-propositions-definitions-objets-textometriques.pdf>

Ken Hyland, Chau Meng Huat, Michael Handford Ken Hyland (2012). *Corpus Applications in Applied Linguistics*. Continuum International Publishing Group.

[2]ANR 07 CORP 030 01 *Rhapsodie, Corpus prosodique de référence en français parlé* (2015) : http://www.projet-rhapsodie.fr/