Du petit fait à la doxa : annotation automatique des anecdotes dans le discours critique sur Molière

Elodie Bénard, Motasem Alrahabi

Du petit fait à la *doxa* : annotation automatique des anecdotes dans le discours critique sur Molière

Bénard, Elodie & Alrahabi, Motasem

Université Paris-Sorbonne – OBVIL
Université Paris-Sorbonne Abu Dhabi
elodie.benard@paris-sorbonne.fr, motasem.alrahabi@gmail.com

1 Anecdotes : rôle, définition et corpus

Nos propres jugements de Molière et nos lectures de son théâtre sont tributaires de la tradition biographique et critique qui s'est constituée après sa mort et qui est caractérisée par la présence massive d'anecdotes. Celles-ci ne sont pas présentes uniquement dans les recueils spécialisés et dans les biographies qui en font leur miel, mais elles se retrouvent dans tous les récits, commentaires et jugements, tant elles sont constitutives du discours sur Molière. Les anecdotes sont constamment répétées, même pour être réfutées. Elles informent en somme toute la réception de l'auteur comique.

En français contemporain, l'anecdote est définie comme le « récit bref d'un petit fait curieux » [Rey, 2005]. Cette acception soulève deux difficultés : le seuil de narrativité et la nature du fait raconté — l'adjectif « curieux » pouvant recouvrir aussi bien les sens « piquant, amusant » que « rare, secret », « intéressant, révélateur ». Une définition fondée sur le seuil de narrativité[1] [K. Abiven, 2015] amène à exclure des séquences peu narrativisées, qui font pourtant allusion à un micro-récit originel, identifiable, et qui attestent son empreinte dans le discours critique. En ce qui concerne la nature du fait rapporté, deux conditions sont nécessaires : le fait ne peut pas être authentifié de façon certaine et il contribue à la construction de l'« écrivain imaginaire », c'est-à-dire, l'écrivain tel qu'il se fait représenter [Diaz, 2007]. Dans notre approche, nous appelons anecdote toute séquence qui a la forme d'un micro-récit ou renvoie sur le mode allusif à un micro-récit, dont la véracité est sujette à caution et qui est révélateur d'une facette de l'instance imaginaire construite par les biographes et les critiques. Cette définition, relativement extensive, n'est pas fondée sur le degré de narrativité et inclut aussi des séquences de longueur variable — les plus courtes contiennent une seule phrase et les plus longues peuvent être composées de plusieurs paragraphes. Elle repose sur trois déterminations essentielles de l'anecdote : sa mise en circulation, sa véridicité problématique et sa fonction symbolique. Ces trois caractéristiques peuvent avoir une traduction formelle, en particulier par des marques linguistiques de la citation de discours ou de la modalisation. Cependant, ces marques peuvent être gommées : « L'une des fonctions les plus fortes de l'idéologie est justement de nous faire croire que tout usage relève d'un usage personnel, d'une appropriation et non d'une réappropriation. La circulation visera donc autant à l'effacement des strates énonciatives qu'à son marquage selon les conditions de circulation du discours. » [Rosier, 2003]. En cela, l'anecdote est proche de discours comme la rumeur. Voici un exemple d'anecdote sur Molière:

> *On rapporte que Molière, en faisant répéter cette pièce, parut mécontent des acteurs qui y jouaient, et principalement de Mlle Beauval, qui représentait le personnage de Toinette. (Corpus Molière – OBVIL)*

Notre corpus est composé de textes écrits après la mort de Molière jusqu'aux premières décennies du XX^{e} siècle : des recueils d'anecdotes du XVIII^{e} siècle, la critique moliéresque produite aux XVIII^{e} et XX^{e} siècles (biographies, ouvrages sur différents aspects de la vie de

Molière et de son théâtre et chapitres d'histoires de la littérature) et les paratextes des éditions des œuvres de Molière. Ces textes, actuellement au nombre de 104, ont été préalablement numérisés dans le cadre du Labex OBVIL de Paris-Sorbonne [Alexandre, 2016] et sont librement accessibles[2]. Eu égard à la taille du corpus (2,5 millions de mots) et à notre objet qui possède des caractéristiques linguistiques identifiables [K. Abiven, 2015], il est pertinent de recourir à l'annotation automatique des textes. Celle-ci permet dans notre contexte de viser un repérage qui soit le plus complet possible, indispensable pour montrer la place occupée par les anecdotes dans le discours critique, mettre en évidence l'invention continuelle de petits récits (favorisée par le caractère répétitif et paradigmatique des schèmes anecdotiques) et repérer les variations formelles et thématiques d'une anecdote au fil du temps.

2 Marqueurs linguistiques des anecdotes

Nous avons isolé un échantillon de textes (≈ 7% de la taille totale du corpus) pour analyser les anecdotes et en identifier les marqueurs. Nous avons alors observé que les anecdotes, en tant que forme de circulation du discours, sont souvent introduites par des indicateurs relevant de la catégorie du médiatif [Guentcheva, 1996], qui permettent à l'énonciateur de se figurer comme passeur d'une information et parfois comme étape dans la chaîne des discours rapportés. Exemples : *au dire de, à en croire, sur la foi de, selon le témoignage de, entendre dire, ouï dire, rapporter, répéter, révéler,* etc.

> *Mais ayant ouï dire que Molière voulait faire une comédie des Femmes savantes, elle supprima sa dissertation. (Corpus Molière – OBVIL)*

Ces indicateurs sont pertinents, mais ne suffisent pas toujours pour identifier les anecdotes. Nous avons donc identifié des indices complémentaires dans le contexte afin de lever l'ambiguïté sur certains indicateurs ou bien d'affiner la catégorisation. Différentes classes d'indices ont ainsi été créées comme les noms propres (*Grimarest, Tallemant*...) et les indices d'encadrement du micro-récit (*un jour, un autre soir*...). Si les signaux de l'ouverture du micro-récit comme « un jour » ou « un soir » sont généralement placés au tout début de la première phrase ou en zone préverbale, la place des marqueurs relevant du médiatif varie.

Les premiers tests ont soulevé une autre difficulté : environ une anecdote sur dix dans le corpus de test ne concerne pas Molière :

> *Racine lui avait confié le rôle d'Oreste ; et ce rôle, suivant une tradition populaire, fut la cause de sa mort ; il [*i. e. *l'acteur Montfleury] se rompit, dit-on, une veine, par les efforts prodigieux qu'il fit pour bien rendre la scène des fureurs. (Corpus Molière – OBVIL)*

Nous avons donc ajouté à la liste des indices complémentaires les variantes du nom « Molière » : *jeune tapissier, poète comédien, patron de la troupe, Poquelin,* etc. Cette première approche va nous aider à mieux cerner les anecdotes sur Molière, mais ne va pas résoudre définitivement le problème, car environ 55% des anecdotes sur Molière dans le corpus analysé ne comportent pas dans leurs phrases introductrices de mention de son nom ou de l'une de ses variantes.

Au total, nous avons collecté pour cette tâche autour de 60 marqueurs linguistiques.

3 Annotation automatique

Pour l'annotation automatique, nous avons utilisé excom2, un outil d'annotation à base de règles et de marqueurs linguistiques de surface [Alrahabi, 2010]. Ceux-ci sont organisés dans

des catégories sémantiques et/ou discursives : *opinions*, *définitions*, *comparaisons*, *conclusions*, *hypothèses*, *expressions ironiques*, etc. La présence dans une phrase d'un indicateur déclenche les règles associées qui explorent le contexte[3] et vérifient la présence ou l'absence d'indices complémentaires [Desclés, 2006]. Voici un exemple de règle :

> **SI** dans une phrase on a un indicateur comme: *en croit, à en croire, s'il faut l'en croire…*
> **ET SI** dans le contexte *Après*, on a un indice comme *un nom propre (Grimarest...), une fonction (auteur...), un terme comme Certains…*
> **ET SI** dans les contextes *Avant* ou *Après* nous n'avons pas de négation comme *ne, pas, rien, plus…*
> **ALORS** annoter la phrase en cours.
> **EXEMPLE** *A en croire Grimarest, l'original de ce maître de philosophie serait Rohault, un des plus zélés et des plus célèbres disciples de Descartes, et en même temps ami de Molière.*

Les règles dans excom2 peuvent être organisées selon un ordre de priorité et utiliser les résultats d'autres règles. Pour les Anecdotes, nous avons créé 5 règles que nous avons associées aux différents marqueurs linguistiques (indicateurs et indices). Une première phase de test sur l'échantillon du corpus était nécessaire pour la stabilisation des règles.

4 Evaluation des résultats

Afin d'évaluer la qualité des annotations automatiques, nous nous sommes focalisés dans un premier temps sur le calcul de la *précision.* Dans cette perspective, nous avons annoté avec excom2 le reste du corpus (≈ 93%) et obtenu 1096 annotations.

Ensuite, nous avons demandé à une personne experte d'évaluer les sorties selon un guide d'annotation. Pour chaque phrase annotée, l'évaluatrice devait choisir entre trois étiquettes que nous présentons dans le tableau suivant avec les résultats de l'évaluation:

Etiquette	**Nombre d'annotations correctes**	**Précision**
Circulation du discours	963	87.9 %
Anecdote	592	54.1 %
Anecdote qui porte sur Molière	425	38.8 %

Vu la particularité du phénomène langagier des anecdotes et la simplicité de notre approche par analyse de surface, nous considérons que ces premiers résultats sont très encourageants et méritent d'être améliorés.

5 Discussion et perspectives

Dans la perspective de l'étude des jugements de valeur sur Molière, les anecdotes sont une voie d'entrée féconde dans le corpus critique car elles éclairent la manière dont s'est élaborée et transformée la *doxa* sur Molière. L'approche que nous avons adoptée nous a permis, avant toute autre chose, de découvrir de nouvelles anecdotes non encore étudiées. Elle nous a fourni une matière abondante et des données quantitatives pour mieux cerner l'objet d'étude.

Nous constatons notamment que les marqueurs linguistiques du récit (comme *un jour*) sont présents dans 20% environ des séquences identifiées comme anecdotes, le reste est repéré grâce à des marqueurs qui renvoient à la circulation du discours et/ou expriment le doute sur sa véridicité. Déterminer précisément, c'est-à-dire quantitativement, le type de marqueurs qui introduit l'anecdote permet de mieux appréhender sa fonction dans le discours biographique et critique en tant qu'élément d'une construction argumentative visant à donner une certaine image de l'auteur et de son œuvre. De plus, les résultats permettent de confronter les séquences que

nous identifions comme anecdotes et les séquences qui présentent des traits linguistiques semblables, mais ne relèvent pas de la définition retenue : là encore, le rapprochement de l'anecdote avec des catégories comme le jugement ou l'opinion aide à mieux cerner la spécificité de l'objet. Afin d'améliorer les résultats, nous allons élargir la couverture des ressources linguistiques, en testant de nouveaux marqueurs comme *dire, déclarer, informer,* etc. ou bien des adjectifs comme *extravagant, inattendu, curieux*, etc. Concernant les anecdotes qui ne sont pas en rapport avec Molière, nous avons constaté que les indices complémentaires que nous avons ajoutés (*Poquelin…*) ne sont pas suffisants : sur la totalité des anecdotes repérées, le système arrive à identifier uniquement 72% d'anecdotes qui concernent Molière. Nous envisageons donc d'explorer d'autres pistes qui vont nous confronter à la problématique du repérage des entités nommées et de l'anaphore.

Les marqueurs des anecdotes dans notre approche permettent de localiser uniquement la phrase qui introduit ou qui conclut l'anecdote, avec, selon le cas, le contenu de l'anecdote. Or, nous avons constaté dans les résultats que la longueur d'anecdote est de 5 à 6 phrases en moyenne. Nous allons donc rechercher des solutions pour délimiter les frontières des anecdotes. Enfin, notre objectif à court terme est de créer et de diffuser un corpus de référence (*Gold Standard Corpus*) où les anecdotes sont manuellement vérifiées et annotées. Ceci nous permettra, entre autres, de mesurer le *rappel* et d'améliorer l'annotation automatique.

Références bibliographiques

Abiven, K. (2015). *L'Anecdote ou la fabrique du petit fait vrai. De Tallemant des Réaux à Voltaire* (1650-1750), Paris, Classiques Garnier, 2015.

Alexandre, D. (2010). Études littéraires et calcul numérique. Présentation, *Revue d'histoire littéraire de la France*, 3, juillet-septembre 2016, Paris. p. 517-520.

Alrahabi, M. (2010). *EXCOM-2: plateforme d'annotation automatique de catégories sémantiques. Applications à la catégorisation des citations en français et en arabe*. Thèse de doctorat, sous la direction du Prof. Jean-Pierre Desclés, Université Paris-Sorbonne.

Bres, J. (1988) À la recherche de la narrativité : éléments pour une théorisation praxématique, dans J. Bres (dir.), *Du récit, encore, Cahiers de praxématique*, Montpellier, Université Paul Valéry-Montpellier III, 11, p. 75-100.

Desclés, J.-P. (2006). *Contextual Exploration Processing for Discourse Automatic Annotations of Texts*, Actes de FLAIRS 2006, Florida, USA

Diaz, J. L. (2007). *L'Écrivain imaginaire. scénographies auctoriales à l'époque romantique*, Paris, H. Champion.

Guentcheva, Z. (1996). *L'Énonciation médiatisée*, Louvain, Peeters, Paris

Rey, A. (2005). *Dictionnaire culturel en langue française*, Paris.

Rosier, L. (2003). *Du discours rapporté à la circulation des discours: l'exemple des dictionnaires de "critique ironique"*, dans Lopez-Muñoz, J.-M., Marnette, S. et L. Rosier (éds). « Formes et stratégies du discours rapporté : Approches linguistique et littéraire des genres de discours », Estudios Lengua y Literatura francesas, 14, p. 63-82.

[1] Seuil en deçà duquel on n'a pas affaire à un récit minimal, défini comme la relation d'actions temporellement ordonnées (t→t+n) et tendue entre un début et une fin permettant de nouer et de dénouer une action [J. Bres, 1988].

[2] http://obvil.paris-sorbonne.fr/corpus/moliere/critique/

[3] Pour la création des espaces de recherche dans un texte, excom2 effectue un prétraitement consistant à segmenter les documents en sections, paragraphes et phrases [Alrahabi, 2010].