HAL
archives-ouvertes.fr

# Annotation manuelle d'expressions référentielles : expérimentations pour simplifier les prises de décisions et optimiser le processus

Frédéric Landragin, Juliette Potier, Meryl Bothua

▶ **To cite this version:**

Frédéric Landragin, Juliette Potier, Meryl Bothua. Annotation manuelle d'expressions référentielles : expérimentations pour simplifier les prises de décisions et optimiser le processus. 9èmes Journées Internationales de la Linguistique de corpus (JLC 2017), Jul 2017, Grenoble, France. halshs-01513810

**HAL Id: halshs-01513810**
**https://halshs.archives-ouvertes.fr/halshs-01513810**

Submitted on 25 Apr 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Annotation manuelle d'expressions référentielles : expérimentations pour simplifier les prises de décisions et optimiser le processus

Landragin, Frédéric, Potier, Juliette & Bothua, Meryl

Laboratoire Lattice
CNRS, ENS, Université de Paris 3, Université Sorbonne Paris Cité, PSL Research University
frederic.landragin@ens.fr

## 1 Introduction

La constitution et l'annotation manuelle d'un corpus regroupant des textes écrits de différents genres textuels et de différentes époques fait partie des objectifs du projet ANR DEMOCRAT[1]. L'objet d'étude est double : les expressions référentielles et les chaînes de référence. La taille envisagée est d'environ un million de mots, soit de l'ordre de 200 000 expressions référentielles annotées, ce qui représente une tâche d'annotation importante. Les arguments sont les suivants : il s'agit premièrement de fournir un corpus de référence qui serve à toute la communauté, deuxièmement de permettre des analyses statistiques (textométriques) avec des données en quantité suffisante, et troisièmement de nourrir un ou plusieurs systèmes d'apprentissage artificiel, de manière à ouvrir la voie de la détection automatique des expressions référentielles et des chaînes de référence en français, suite aux expérimentations déjà réalisées avec le corpus ANCOR (Désoyer *et al.*, 2014), seul corpus de taille comparable (115 000 anaphores annotées).

Compte tenu de la taille de corpus envisagé, la procédure d'annotation manuelle doit être la plus efficace possible, afin de ne pas faire appel à des moyens humains démesurés. Nous décrivons dans cette présentation les discussions et les expérimentations qui ont permis de définir une procédure d'annotation rationnelle pour l'annotation des mentions, c'est-à-dire pour la phase la plus coûteuse de l'annotation – et celle sur laquelle reposeront les phases suivantes, comme celle d'annotation des chaînes de référence. Nos choix reposent sur l'expérience des annotateurs du corpus ANCOR (Muzerelle *et al.*, 2014) et du corpus MC4 (Landragin, 2011), sur des calculs d'accords inter-annotateurs et sur des expérimentations chronométrées comparatives.

## 2 La référence et les expressions référentielles

### 2.1 Annoter quoi et comment ?

La référence est un objet linguistique très vaste, qui a fait l'objet de très nombreuses publications (Charolles 2002). Le projet MC4 (Landragin, 2011) s'était intéressé aux multiples facteurs morphologiques, syntaxiques, sémantiques et pragmatiques qui interviennent lors de la résolution des références, c'est-à-dire l'attribution d'un référent à une expression référentielle, en incluant l'attribution d'un antécédent à une anaphore. La procédure d'annotation résultante avait impliqué l'annotation manuelle de ces facteurs, du moins d'une sélection d'une dizaine de facteurs considérés comme déterminants. Au final, le corpus n'a pas dépassé 5 000 expressions référentielles annotées. Pour le projet DEMOCRAT, il n'était pas question de reproduire une procédure aussi détaillée, et ce d'autant plus que l'annotation de certains de ces facteurs est

automatisable. Nous avons ainsi choisi d'annoter uniquement le résultat de la résolution de la référence. Deux possibilités apparaissent ici : soit on saisit, pour chaque expression, un identifiant du référent ; soit on regroupe les expressions en chaînes (anaphoriques et/ou coréférentielles), ce qui fait l'économie des identifiants des référents mais nécessite de construire des chaînes, autrement dit des objets non liés à un et un seul marquable.

Une première expérimentation a permis de comparer les deux méthodes et a soulevé l'importance décisive de l'ergonomie de l'outil d'annotation utilisé : comme il est possible de déduire automatiquement les chaînes à partir d'une annotation des mentions en identifiants, seule compte la rapidité d'action. Or, quand l'outil est bien choisi et permet la complétion automatique de l'identifiant en cours de saisie, il s'avère que la méthode à base d'identifiants est plus rapide que celle à base de construction de chaînes. En effet, manipuler un objet couvrant potentiellement le texte entier est bien plus délicat et propice à des erreurs que saisir des identifiants localement, au niveau du marquable qu'est l'expression référentielle. Plusieurs outils ont été testés (MMAX2, GLOZZ, ANALEC) et c'est finalement l'implémentation de la complétion dans l'outil ANALEC qui a permis la plus grande efficacité (Landragin *et al.*, 2012).

### 2.2 L'accord inter-annotateurs pour la référence

S'il n'est pas possible de valider une procédure d'annotation (à moins de disposer d'un *gold standard*), il est possible d'évaluer sa reproductibilité, ce qui donne une indication précieuse sur l'intérêt des annotations d'un corpus (Mathet et Widlöcher, 2016). Ceci se fait en impliquant plusieurs annotateurs – 2 à 3 dans nos expérimentations – puis en calculant l'accord inter-annotateurs. Plusieurs indicateurs statistiques sont couramment utilisés : $\alpha$, $\pi$, $\kappa$ et désormais $\gamma$ (Mathet *et al.*, 2015). Même si nous avons calculé lors de chaque comparaison importante de procédures l'ensemble des indicateurs, nous soulignons qu'aucune expérimentation n'a permis d'obtenir des scores très élevés : autour de 0,53 pour $\gamma$ dans les premières expérimentations, autour de 0,73 dans les dernières ($\alpha$ entre 0,662 et 0,733, $\pi$ entre 0,66 et 0,732, $\kappa$ entre 0,661 et 0,733). Ceci est dû à la nature de l'objet d'étude : la coréférence et l'anaphore sont des phénomènes complexes, pour lesquels les interprétations peuvent varier d'un annotateur à l'autre sans pour autant que les annotations en deviennent inutisables. Les corpus constitués dans d'autres langues que le français font face à des scores similaires aux nôtres, et la communauté s'est habituée à des taux d'accord modestes (Artstein et Poesio, 2008).

## 3 Identifier des stratégies d'annotation

Nos premières expérimentations d'annotation se sont déroulées sur des textes narratifs, en l'occurrence des extraits de romans libres et gratuits, disponibles sur la plateforme *wikisource*. Il s'agit donc de textes littéraires, et les discussions ont vite porté sur le filtrage ou non des référents : faut-il annoter toutes les expressions référentielles, ou seulement celles qui réfèrent à des personnages humains, à des êtres animés, à des objets concrets, etc. Nous constatons par exemple que les expressions référentielles temporelles sont très peu reprises et ne forment donc que rarement des chaînes de référence intéressantes à étudier. Dans ce cas, faire l'impasse sur l'annotation de ces « singletons » permettrait d'aller plus vite à l'essentiel.

Sauf que : 1. l'annotation systématique de toutes les expressions référentielles permet de nourrir un système d'apprentissage dédié à la détection des expressions référentielles (ce qui, en TAL, est une tâche très complexe, différente de celles consistant à détecter les entités nommées et les pronoms anaphoriques) ; 2. il n'est pas possible de savoir si l'expression en cours d'annotation

va être reprise ultérieurement ou non (autrement dit la tâche peut comporter des retours en arrière dans le texte si l'annotateur s'aperçoit qu'il a oublié une expression – initialement considérée comme singleton) ; 3. se demander à chaque expression si elle a des chances d'être un singleton ou de faire partie d'une chaîne de référence va à l'encontre de l'aspect « robotique » et efficace de l'annotation : se poser trop de questions est parfois contre-productif, et il vaut mieux tout annoter en se posant moins de questions. Nous avons convenu de rendre le choix des identifiants de référents le plus rapide possible pour les expressions peu susceptibles d'être reprises. Pour les expressions qui resteront clairement des singletons, nous employons un code dédié (« SI » comme singleton), ce qui permet d'augmenter encore l'efficacité.

## 4 Pré-annotation : aide ou gêne ?

L'aspect robotique du repérage des expressions référentielles (pas de leur attribution d'un référent) peut être encouragé en utilisant un système de TAL en tant que pré-annotateur. Mais encore faut-il trouver un trouver un système de TAL qui soit adapté au français et dont le taux d'erreur ne soit pas une entrave à l'annotation : quand les erreurs sont nombreuses, l'annotateur a vite l'impression de passer son temps à les corriger plutôt qu'à exploiter directement les pré-annotations. Or il n'existe pas de système de détection automatique des expressions référentielles, et il nous faut nous rabattre sur la détection des entités nommées (ce qui est une tâche plus réduite, cf. Nouvel *et al*., 2015), la détection des anaphores (ce qui est également très réducteur) ou bien sûr la détection des chaînes de référence – sauf que le seul système disponible pour le français, RefGen (Todiraşcu et Longo, 2011) a des performances moyennes. En fin de compte, le système le plus proche du résultat souhaité est tout simplement un détecteur de *chunks* nominaux. Après un comparatif des performances des différents outils disponibles, notre choix a porté sur le *chunker* nominal de SEM (Tellier *et al.*, 2012).

Le principal avantage d'un *chunker* nominal est qu'il permet à l'annotateur de n'oublier aucune expression référentielle. Mais quelques inconvénients viennent contrebalancer cet avantage : 1. tous les *chunks* nominaux d'un texte ne réfèrent pas, donc le *chunker* produit du bruit qui peut perturber l'annotateur (« il » impersonnel, mention non référentielle de partie du corps comme « avoir la grosse tête », etc.) ; 2. un *chunker*, par définition, repère des portions de texte non enchâssées – or des expressions référentielles peuvent s'enchâsser, comme les compléments du nom. Des adaptations des résultats du *chunker* sont donc nécessaires et, là aussi, tout repose sur l'ergonomie de l'outil d'annotation utilisé. Ainsi, un outil qui permet de rectifier facilement les frontières d'un marquable peut avantager l'exploitation d'une pré-annotation en *chunks*.

## 5 Conclusion

Les objectifs textométriques et TAL de notre corpus encouragent à privilégier certaines stratégies d'annotation par rapport à d'autres. L'intérêt de disposer d'une annotation – même minime – de toutes les expressions référentielles nous incite à diriger la procédure d'annotation dans ce sens. C'est une décision qui privilégie la référence à la coréférence, mais plusieurs expérimentations chronométrées (tranches de 30 minutes d'annotation avec des conditions différentes) ont montré que le surcoût en temps d'annotation reste raisonnable compte tenu du gain final : dans le pire des cas (sur 7 chronométrages), la perte de temps est de 50%. Quant à l'utilisation d'un *chunker* pour disposer d'un pré-repérage des expressions référentielles, elle a été laissée au choix de l'annotateur : comme beaucoup d'aspects de la tâche d'annotation, l'appropriation de la procédure se fait mieux si une certaine souplesse est autorisée, sous condition bien entendu que les annotations finales se soient pas impactées, ce qui est le cas ici.

## Références bibliographiques


Artstein, R., Poesio, M. (2008). Inter-Coder agreement for Computational Linguistics, *Computational Linguistics, 34*, 555-596.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics, 22*, 249-254.

Charolles, M. (2002). *La référence et les expressions référentielles en français.* Paris : Ophrys.

Désoyer, A., Landragin, F., Tellier, I., Lefeuvre, A., Antoine, J.-Y. (2014). Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ANCOR. *Traitement Automatique des Langues, 55(2)*, 97-121.

Heiden, S., Magué, J.-P., Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In: *Proceedings of Tenth International Conference on the Statistical Analysis of Textual Data*, Vol. 2, 1021-1032.

Krippendorff, K. (2012). *Content analysis: an introduction to its methodology (third edition).* Thousand Oaks : Sage Publishing.

Landragin, F. (2011). Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits. *Corpus, 10*, 61-80.

Landragin, F., Poibeau, T., Victorri, B. (2012). ANALEC: a New Tool for the Dynamic Annotation of Textual Data. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, Istanbul, Turquey, 357-362.

Mathet, Y., Widlöcher, A. (2016). Évaluation des annotations : ses principes et ses pièges. *Traitement Automatique des Langues, 57(2)*, 73-98.

Mathet, Y., Widlöcher, A., Métivier, J.-P. (2015). The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment. *Computational Linguistics, 41(3)*, 437-479.

Müller, C., Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In: Braun, S., Kohn, K., Mukherjee, J. (Eds.). *Corpus technology and language pedagogy: New resources, new tools, new methods.* Frankfurt : Peter Lang.

Muzerelle, J., Lefeuvre, A., Schang, E., Antoine, J.-Y., Pelletier, A., Maurel, D., Eshkol, I., Villaneau, J. (2014). ANCOR CENTRE, a large free spoken french coreference corpus: description of the resource and reliability measures. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Nouvel, D., Ehrmann, M., Rosset, S. (2015). *Les entités nommées pour le traitement automatique des langues.* Londres : Éditions ISTE.

Tellier, I., Duchier, D., Eshkol, I., Courmet, A., Martinet, M. (2012). Apprentissage automatique d'un chunker pour le français. In *Actes de la 19e Conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble.

Todiraşcu, A., Longo, L. (2011). RefGen, outil d'identification automatique des chaînes de référence en français. 18e Conférence sur le Traitement Automatique des Langues Naturelles, session des démonstrations industrielles, Montpellier.

Widlöcher, A., Mathet, Y. (2009). La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. In : *Actes de la 16e Conférence sur le Traitement Automatique des Langues Naturelles*, Senlis.


---

[1] DEMOCRAT, « description et modélisation des chaînes de référence : outils pour l'annotation de corpus (en diachronie et en langues comparées) et le traitement automatique », projet ANR-15-CE38-0008.