# *RefGen*, outil d'identification automatique des chaînes de référence en français

Laurence Longo  Amalia Todirascu

Université de Strasbourg, 22 avenue René Descartes, 67084 Strasbourg Cedex, France
longo@unistra.fr, todiras@unistra.fr

Nous présentons *RefGen*, un outil d'identification automatique des chaînes de référence (CR) en français. Les CR sont composées d'au moins trois expressions référentielles (Schnedecker, 1997). Développé dans un cadre industriel[1], *RefGen* est un prototype (développé en Perl et en Java) pouvant être intégré dans un système de détection automatique de thèmes. L'architecture de *RefGen* est modulaire et composée d'un étiquetage fin, d'un module d'annotation des expressions référentielles (groupes nominaux simples et complexes, entités nommées) et d'un module de calcul de la référence. *RefGen* utilise aussi une série de paramètres spécifiques au genre textuel pour calculer les relations de référence (distance entre les maillons d'une CR, nombre de maillons d'une CR, etc).

Pour l'étiquetage, *RefGen* utilise le catégoriseur TTL[2] (Ion, 2007) dans sa version française. Développé en Perl, TTL utilise le jeu d'étiquettes morphosyntaxiques fin proposé dans le projet Multext (Ide et Véronis, 1994), permettant de préciser des informations comme le genre, le nombre, le mode, le temps, etc. En plus de cet étiquetage fin, TTL identifie certains noms propres et fournit une analyse syntaxique partielle en chunks (groupes nominaux, groupes prépositionnels, groupes adjectivaux et groupes verbaux). Les sorties étiquetées sont disponibles en format XML.

Pour identifier les relations de référence entre les différentes entités du discours, *RefGen* annote d'abord les diverses expressions référentielles contenues dans les CR (groupes nominaux, entités nommées) avec RefAnnot. Ainsi, le module d'annotations RefAnnot (développé en Java) applique un ensemble de règles morphosyntaxiques pour identifier les expressions référentielles. Ces règles sont définies dans un format XML facilement transformable vers un autre format. RefAnnot identifie les groupes nominaux complexes (CNp, groupes nominaux modifiés par deux groupes prépositionnels au plus), plus informatifs, qui introduisent de nouveaux éléments dans le discours ainsi que certaines entités nommées (noms de personnes, organisations, lieux et fonctions). Pour faciliter le tri des divers candidats anaphoriques, les emplois impersonnels du pronom « il » sont aussi annotés. Ainsi, *RefGen* ne les prendra pas en compte lors de son calcul. Les sorties de RefAnnot sont disponibles en format xml mais aussi en html (pour faciliter la lisibilité).

Le texte enrichi en annotations passe alors dans le module de calcul de la référence CalcRef (développé en Java). L'algorithme mis en place utilise des paramètres liés au genre textuel pour sélectionner les premiers maillons des chaînes de référence mais aussi un score d'accessibilité global calculé à partir de l'échelle d'accessibilité d'(Ariel, 1990). Puis, la sélection des paires antécédent-anaphore s'effectue par la validation d'une série de contraintes (lexicales, syntaxiques, sémantiques) fortes et faibles. Une fois les paires identifiées, *RefGen* construit les CR suivant la propriété de transitivité.

Outre son utilisation première (l'identification des CR), *RefGen* peut aussi être utilisé comme outil de pré-annotation de corpus. D'autres règles morphosyntaxiques peuvent être facilement ajoutées à RefAnnot suivant les besoins (identification de dates, d'évènements, ou typage plus fin des entités nommées (fonction administrative, fonction politique, etc.). Une interface graphique est en cours de réalisation pour faciliter l'utilisation de l'outil.

ARIEL M. (1990). *Accessing Noun-Phrase Antecedents*. Londres : Routledge.

IDE N., VERONIS J. (1994). MULTEXT (Multilingual Tools and Corpora). *Actes de IAACL*, Kyoto.

ION R. (2007). Word Sense Disambiguation Method Applied to English and Romanian. Thèse de doctorat, Bucharest.

SCHNEDECKER C. (1997). « Nom propre et chaînes de référence », *Recherches Linguistiques*, 21, Paris, Klincksiek.

[1] L'outil a été développé dans le cadre d'une convention CIFRE avec la société RBS, Strasbourg (www.rbs.fr).

[2] TTL est disponible comme service Web sur la plate-forme Weblicht (https://weblicht.sfs.uni-tuebingen.de/). Un code d'accès est nécessaire (disponible sur simple demande).