RefGen : un module d'identification des chaînes de référence dépendant du genre textuel

Laurence Longo Amalia Todiraşcu

Laboratoire LiLPa, EA 1339, Université de Strasbourg, 67000 Strasbourg Cedex, France
longo@unistra.fr, todiras@unistra.fr

Résumé

Dans cet article, nous présentons *RefGen*, un module d'identification des chaînes de référence pour le français. *RefGen* effectue une annotation automatique des expressions référentielles puis identifie les relations de coréférence établies entre ces expressions pour former des chaînes de référence. Le calcul de la référence utilise des propriétés des chaînes de référence dépendantes du genre textuel, l'échelle d'accessibilité d'(Ariel, 1990) et une série de filtres lexicaux, morphosyntaxiques et sémantiques. Nous évaluons les premiers résultats de *RefGen* sur un corpus issu de rapports publics.

Abstract

We present *RefGen*, a reference chain identification module for French. *RefGen* automatically annotates referential expressions then identifies coreference relations between these expressions to make reference chains. Reference calculus uses textual genre specific properties of reference chains, (Ariel, 1990)'s accessibility theory and applies lexical, morphosyntactic and semantic filters. We evaluate the first results obtained by *RefGen* from a public reports corpus.

Mots-clés : Chaînes de référence, relation de coréférence, saillance, genre textuel
Keywords: Reference chain identification, coreference relation, salience, genre

1 Contexte et motivation

L'identification automatique des chaînes de référence est un problème difficile, vu la variété des expressions linguistiques intervenant dans leur construction et la nécessité d'en détenir des connaissances linguistiques et extra-linguistiques approfondies. Les expressions référentielles (noms propres, descriptions définies, démonstratifs, pronoms) ont fait l'objet de nombreuses études qui se sont focalisées sur leur description linguistique (Kleiber, 1994), ou sur leur rôle dans l'organisation du texte (Charolles, 1997).

Les chaînes de référence constituent des marques linguistiques de la cohérence du texte permettant d'identifier la continuité ou la rupture thématique. De plus, (Schnedecker, 2005) montre que la structure des chaînes de référence et le choix des expressions référentielles qui les constituent sont dépendants du

genre textuel. Pour une identification automatique de ces chaînes, plusieurs travaux proposent des modèles cognitifs pour le calcul de la coréférence (Salmon-Alt, 2001) ou exploitent des indices linguistiques de surface (Hernandez, 2006), mais peu de modèles opérationnels rendent compte des propriétés des chaînes de référence suivant le genre textuel. L'identification des chaînes de référence nécessite d'abord l'identification des relations de coréférence entre les expressions anaphoriques et leurs antécédents. Ensuite, la construction des chaînes de référence s'appuie sur la transitivité de la relation de coréférence. Les méthodes de résolution de la coréférence appliquent soit des règles de sélection d'antécédents définies manuellement, soit des règles apprises à partir de corpus annotés. Malgré l'existence de quelques corpus français annotés en coréférence (Manuélian, 2003), (Salmon-Alt, 2001), il est difficile d'utiliser ces ressources pour l'apprentissage automatique, vu la taille réduite et l'hétérogénéité des phénomènes annotés (pronoms personnels, descriptions définies, coréférence, anaphores associatives)[1].

De ce fait, notre travail se situe dans la lignée des méthodes symboliques (Hartrumpf, 2001), (Bontcheva et al, 2002), (Mitkov, 2001) qui se montrent plus facilement adaptables aux nouveaux domaines ou applications. Pour ces systèmes, la relation de coréférence est établie en appliquant des règles qui repèrent les antécédents possibles, après vérification de critères de compatibilité des propriétés (morphosyntaxiques, syntaxiques, sémantiques) des candidats. Pour chaque candidat anaphorique, le choix des antécédents possibles s'effectue selon un calcul de saillance (Victorri, 2005), la vérification des propriétés de cohérence locale et globale du discours, l'application de contraintes (Beaver, 2004) ou un calcul d'accessibilité des expressions référentielles (Ariel, 1990). Notre module d'identification des chaînes de référence *RefGen* applique une nouvelle méthode de calcul de la référence qui, en plus de l'accessibilité et de la vérification de contraintes, applique des paramètres dépendants du genre textuel (préférence pour une catégorie d'expressions référentielles, distance moyenne entre les maillons). Développé dans un cadre industriel[2], *RefGen* est intégré à un outil de détection de thèmes qui utilise des marqueurs linguistiques (les chaînes de référence) pour optimiser les résultats des moteurs de recherche et la navigation dans les documents.

Dans cet article, nous présentons les propriétés des chaînes de référence et précisons leurs caractéristiques pour le genre textuel de rapports publics. Nous nous focalisons ensuite sur les prétraitements mis en place (étiquetage, identification des expressions référentielles) pour *RefGen* et présentons l'algorithme d'identification des chaînes de référence. Nous évaluons *RefGen* sur un corpus de rapports publics français et montrons la plus-value apportée par la prise en compte des paramètres spécifiques au genre.

2 Les chaînes de référence

En suivant (Schnedecker, 2005), nous considérons qu'une chaîne de référence est une relation qui s'établit entre trois maillons au moins (trois expressions coréférentielles). Ainsi, l'exemple suivant compte une seule chaîne de référence composée de trois maillons (en gras) : ***M. Pons*** *rappelle que Jacques Chirac* ***lui*** *apparaît comme « le candidat légitime » de* ***son*** *parti. (Le Monde Diplomatique, 1980-1988).*

Les chaînes de référence comprennent trois types de constituants à fonction référentielle : les noms propres (Np), les groupes nominaux (défini, indéfini, possessif ou démonstratif) et les pronoms. Les Np jouent un rôle important dans la structuration du discours car ils se trouvent souvent en tête d'une chaîne de référence dans les portraits journalistiques (Schnedecker, 2005). En dehors des cas de compétition référentielle où la répétition du Np élimine une ambigüité entre deux référents, la redénomination d'un Np marque une rupture dans la chaîne (un changement d'acteur). Lorsqu'une expression référentielle est utilisée, elle déclenche un « processus de recrutement » particulier du référent. De ce fait, le démonstratif (e.g. *ce*

[1] La campagne SemEval 2010 *Coreference Resolution in Multiple Languages* ne propose pas de données pour le français.
[2] Le projet de détection automatique de thèmes s'effectue en collaboration avec l'entreprise RBS (Strasbourg).

président) renvoie directement au référent sur la base d'un critère de proximité alors que le pronom « il » indique de recruter un référent qui soit l'argument d'une proposition saillante (Kleiber, 1994). Ainsi, un indice est donné quant au référent à « garder en mémoire » et qui constitue alors un thème local dans cette portion du discours. En revanche, l'emploi d'expressions nominales alors que le contexte ne le nécessite pas (e.g. lorsqu'un pronom aurait suffit), constitue un indice de rupture. Ces indices seront exploités par le système de détection de thèmes.

Dans notre approche, nous travaillons sur les relations mono référentielles (à l'exclusion d'anaphores plurielles) s'établissant entre des expressions co-référentes inter et intra phrases. Aussi, sont considérées les situations de coréférence directe (Manuélian, 2002) où les groupes nominaux coréférents possèdent la même tête nominale (e.g. « *le changement climatique* / *ce changement* ») ainsi que les relations de coréférence entre les noms de personne et les noms de fonction[3]. Nous faisons l'hypothèse que les chaînes de référence possèdent des caractéristiques dépendantes du genre textuel (Longo, Todirașcu, 2010). En suivant la typologie de (Schnedecker, 2005), nous avons comparé les chaînes de référence présentes dans un corpus de genres divers (journaux, éditoriaux, rapports publics, romans, normes européennes) de 50 000 mots[4]. Ainsi, cinq critères de comparaison ont été établis et appliqués pour chaque genre étudié : la longueur de la chaîne, la distance moyenne entre les maillons, la catégorie du premier maillon, la catégorie la plus fréquente des maillons, la position thématique. Pour les rapports publics, nous avons remarqué que la longueur moyenne des chaînes était de 3-4 maillons, alors qu'elle était en moyenne trois fois plus élevée pour un roman. La distance moyenne entre les maillons des rapports publics est de deux phrases, mais seulement d'une phrase pour le roman, tant les réflexifs et possessifs sont utilisés dans ce dernier genre. Aussi, de même que (Schnedecker, 2005), nous avons constaté que la catégorie des noms propres était privilégiée dans les premiers maillons des chaînes des textes journalistiques (39,5%) alors qu'il s'agissait plutôt des descriptions définies pour les rapports publics (45%). Ce sont aussi les descriptions définies qui représentent un tiers des maillons dans les rapports publics alors qu'il s'agit des pronoms (50%) dans le roman. Enfin, un dernier critère utile pour identifier le thème du document, le premier maillon d'une chaîne, coïncide peu avec le thème phrastique (40%) pour les rapports publics mais très souvent (80%) pour les textes journalistiques. A l'issue de cette étude, nous avons dégagé des propriétés des chaînes de référence issues d'un genre textuel particulier. Nous présentons dans les sections suivantes le module *RefGen* qui tient compte de ces propriétés.

3 Annotations des expressions référentielles dans *RefGen*

Pour identifier les expressions référentielles, le texte est tout d'abord étiqueté par TTL (Ion, 2007). En plus de la segmentation en chunks simples (groupes nominaux (Np), groupes prépositionnels (Pp), groupes adjectivaux (Ap)), cet étiqueteur fournit des informations morpho-syntaxiques (temps, mode, personne, genre, nombre). Partant de la sortie XML, nous appliquons une base de patrons symboliques pour identifier des expressions plus complexes susceptibles d'être présentes au début des chaînes de référence (car plus informatives) : les groupes nominaux complexes (CNp)[5] (e.g. *l'élévation du niveau moyen global de la mer)* et les entités nommées (nom de fonction, d'organisation ou de personne). Les emplois impersonnels du pronom « *il* » sont aussi annotés (e.g. « *il pleut* »), afin d'ignorer ces emplois dans le calcul des chaînes. Voici un exemple d'annotations comprenant les lemmes, chunks simples (non-récursifs) et complexes

[3] Nous ne traitons que quelques cas de coréférence indirecte parce que nous ne faisons pas appel à des connaissances externes.

[4] Nous avons travaillé sur des extraits de notre corpus de départ (de 500 000 tokens), tant l'annotation manuelle des chaînes de référence est une activité fastidieuse. Nous comptons poursuivre l'annotation de ce corpus de référence, vu qu'il n'en existe pas encore pour le français.

[5] Un CNp est un groupe nominal modifié par deux groupes prépositionnels au plus, ou bien un groupe nominal modifié par une proposition relative (une proposition simple contenant un prédicat et un complément d'objet direct ou indirect).

(CNp), propriétés morpho-syntaxiques (ana), entités nommées (NER) et *il* impersonnel (imp) (voir Figure 1). Nous exploitons ces annotations automatiques pour construire les chaînes de référence.

```
<w lemma="le" chunk="Np#1" ana="Da-fs">L'</w>
<w lemma="union" chunk="Np#1" ana="Ncfs" ner="NER#1, org">Union</w>
<w lemma="européen" chunk="Np#1, Ap#1" ana="Af-fs" ner="NER#1, org">européenne</w>
<w lemma="avoir" chunk="Vp#1" ana="Vaip3s">a</w>
<w lemma="adopter" chunk="Vp#1" ana="Vmps-s">adopté</w>
<w lemma="il" ana="Pp3ms" feat="imp">il</w>
<w lemma="y" ana="Pp3">y</w>
<w lemma="avoir" ana="Vaip3s">a</w>
<w lemma="peu" chunk="Ap#2" ana="R">peu</w>
<w lemma="de_le" chunk="CNp#5, Pp#1, Np#2" ana="Dg-mp">des</w>
<w lemma="acte" chunk="CNp#5, Pp#1, Np#2" ana="Ncmp">actes</w>
<w lemma="législatif" chunk="CNp#5, Pp#1, Np#2, Ap#3" pana="Af-mp">législatifs</w>
<w lemma="relatif" chunk="CNp#5, Pp#1, Np#2, Ap#3" ana="Af-mp">relatifs</w>
<w lemma="à+le" chunk="CNp#5, Pp#2, Np#3" ana="Dg-ms">au</w>
<w lemma="changement" chunk="CNp#5, Pp#2, Np#3" ana="Ncms">changement</w>
<w lemma="climatique" chunk="CNp#5, Pp#2, Np#3, Ap#4" ana="Af-ms">climatique</w>
```

Figure 1 : exemple de sortie enrichie en annotations dans *RefGen*

4 Algorithme d'identification des chaînes de référence dans *RefGen*

Le texte enrichi en annotations (section 3) est traité pour extraire les chaînes de référence. Connaissant le genre du texte (le genre doit être précisé par l'utilisateur), *RefGen* va alors sélectionner préférentiellement les candidats correspondant aux paramètres identifiés dans (Longo et Todirascu, 2010). Pour les rapports publics, *RefGen* va ainsi privilégier les descriptions définies comme premiers maillons, les chaînes de référence courtes (3 à 4 maillons) et les antécédents situés à une distance moyenne de 2 phrases. Le calcul de la référence (CalcRef) s'effectue en plusieurs étapes.

Sélection du premier maillon. La sélection des candidats susceptibles de constituer le premier maillon de la chaîne dépend de l'accessibilité globale (AG) du candidat, de la fonction syntaxique (la fonction sujet est privilégiée) et du type privilégié (pour le 1er maillon) dépendant du genre. En effet, sont sélectionnés les candidats ayant l'AG la plus élevée : les noms propres ou les CNp. L'AG est calculée à partir de l'échelle d'accessibilité d'(Ariel, 1990) qui classe les expressions référentielles suivant leur degré d'informativité, de rigidité et d'atténuation. Pour chacun des degrés, nous attribuons un score de 10 à 110. Ainsi, l'AG d'une expression référentielle constitue la somme des trois scores. Par exemple, l'AG d'un nom propre complet ("*le président Barack Obama*") est de 220 (100+100+20), l'AG du pronom *elle* est de 150. La fonction syntaxique est attribuée à l'aide de règles heuristiques qui déterminent la position du candidat par rapport au verbe principal (sujet, objet direct, objet indirect, complément du nom). Un poids supplémentaire à l'AG est alors affecté pour la fonction (e.g. 100 pour la position sujet, 50 pour la position d'objet) et pour le type (50 si type privilégié). Le candidat ayant le score total le plus élevé est sélectionné.

Sélection des autres maillons. Ensuite, pour déterminer les autres maillons des chaînes de référence (rangs 2, 3 et suivants), nous recherchons des liens de coréférence s'établissant entre les candidats d'accessibilité haute (les premiers maillons potentiels) et ceux d'accessibilité basse (les autres maillons candidats). Nous procédons à une identification des paires antécédent-anaphore présentes dans un intervalle correspondant à la distance moyenne entre les maillons (paramètre dépendant du genre). Puis, pour départager les antécédents potentiels, nous adaptons l'approche de (Gegg-Harrison et Byron, 2004) qui ont proposé une liste de contraintes (morpho-syntaxiques, sémantiques) à satisfaire pour chaque paire de candidats antécédent-anaphore. Nous définissons deux types de filtres à appliquer : des filtres forts et

des filtres faibles. Les filtres forts permettent d'éliminer certaines paires impossibles. Néanmoins, ces filtres forts ne s'appliquent pas à certaines anaphores, comme les pronoms réfléchis. Par exemple, le filtre fort d'imbrication permet d'éliminer des paires où les deux candidats sont imbriqués (e.g. une description définie et son complément de nom : *«[les répercussions [du changement climatique]] »*) car ils ne peuvent pas être coréférents. Les filtres faibles permettent d'identifier les candidats qui valident le plus de filtres. On vérifie par exemple la compatibilité en termes de fonction syntaxique ou en genre et nombre entre un candidat nominal et une anaphore pronominale. Une fois les relations anaphoriques établies, nous regroupons les anaphores ayant un antécédent commun dans une même chaîne de référence.

Exemple. Nous appliquons l'algorithme suivant les paramètres spécifiques au genre de rapports publics :

Pour soutenir [la coopération en matière d'adaptation]**i** et guider [les progrès du cadre d'action européen]**j**, [la Commission]**k** a [l'intention]**l** de créer [un groupe de pilotage consacré [aux incidences du changement climatique]**p**]**m**. [Ce groupe]**r** réunira [des représentants des États membres de l'UE]**s** qui participent [à la formulation de programmes d'adaptation nationaux]**t** et [il]**w** consultera [des représentants de la communauté scientifique]**y**.

Après le calcul de saillance des candidats, nous obtenons 5 premiers maillons potentiels : i, j, m, s et y. Ensuite, nous générons toutes les paires antécédent-anaphore possibles. Certaines paires sont éliminées dès le début, comme par exemple (i, r) car la tête lexicale est différente, ou (m, p) car les groupes sont imbriqués. Pour le candidat *il*, parmi les paires (i, *il*) (j, *il*), (k, *il*), (l, *il*), (m, *il*)…, le maximum de filtres satisfaits est réalisé pour [ce groupe]**r**. De plus, comme (m, r) est une paire valide (tête lexicale identique), nous obtenons ainsi une chaîne possible : {un groupe de pilotage […], ce groupe, il}.

5 Evaluation du module *RefGen*

Nous avons évalué *RefGen* en comparant les annotations (entités nommées, groupes nominaux complexes et *il* impersonnel) et le calcul de la référence (CalcRef) obtenus par rapport à une annotation manuelle. Le corpus d'évaluation est composé de rapports publics issus de la Commission Européenne. L'extrait choisi porte sur le changement climatique et compte 7230 mots. Il contient 118 relations anaphoriques (impliquant surtout des anaphores nominales, démonstratives ou définies) et 24 chaînes de référence. Nous avons reporté les mesures de rappel, précision et performance (*f-mesure*) (Tableau 1) pour chacune des annotations et le calcul de la référence pour les paires antécédent-anaphore et les chaînes retrouvées.

	Ner	CNp	Il imp	CalcRef (paires)	CalcRef (chaînes)
rappel	0,85	0,87	0,91	0,69	0,58
précision	0,91	0,91	1	0,78	0,70
f-mesure	0,88	0,89	0,95	0,73	0,63

Tableau 1 : Evaluation de *RefGen*

Pour les Ner, les erreurs ont porté essentiellement sur des abréviations de CNp (e.g. *GES : gaz à effet de serre*) qui ont été considérées comme des sigles et qui ont de ce fait été étiquetées <Org>. Les erreurs de CNp sont dues à une sous-spécification (les patrons ne prenaient pas en compte les groupes nominaux modifiés par plus de deux groupes prépositionnels). Du côté des *il* impersonnels, nous observons de bonnes performances car le corpus contenait peu d'occurrences ambigües. Concernant CalcRef, les erreurs pour les paires résultent de problèmes d'étiquetage (erreur dans le genre/nombre) ou parce que plusieurs antécédents d'un même candidat ont satisfait le même nombre de filtres mais qu'ils n'étaient pas coréférents. Les erreurs d'identification des chaînes sont par conséquent liées aux fausses paires antécédent-anaphore retrouvées ainsi qu'aux erreurs d'étiquetage (sigles). Nous avons mis en place des patrons de correction pour les erreurs récurrentes de TTL. Aussi, nous pensons rajouter des listes de synonymes et des relations ontologiques qui permettraient d'éliminer certaines paires antécédent-anaphore.

Nous avons ensuite testé l'importance de la prise en compte des paramètres liés au genre textuel dans CalcRef. Si nous appliquons les paramètres du genre journalistique à notre corpus d'évaluation, nous obtenons une performance similaire pour les paires (f-mesure=0,7) mais une baisse significative pour les chaînes de référence (f-mesure=0,54). Sans aucun paramètre (la longueur des chaînes n'est pas limitée ni la distance entre les maillons), nous obtenons une performance comparable pour les paires (f-mesure=0,71) mais une baisse importante de la f-mesure pour la détection des chaînes (0,51). Le nombre de paires est sensiblement le même (quelques paires sont rajoutées entre des candidats éloignés). En revanche, parce que certaines chaînes identifiées regroupent des chaînes plus courtes, nous obtenons peu de chaînes valides.

6 Conclusion et perspectives

Dans cet article, nous avons présenté le module *RefGen* qui identifie automatiquement les chaînes de référence à partir d'indices linguistiques de surface, mais de peu de connaissances externes. Notre module emploie divers prétraitements linguistiques pour annoter les expressions référentielles contenues dans les chaînes de référence. Pour le calcul de la référence, *RefGen* utilise des paramètres liés au genre textuel et une série de filtres forts et faibles. L'évaluation est encourageante et permet d'apprécier la plus-value apportée par la prise en compte des paramètres liés au genre. Nous poursuivons notre travail en ce sens sur des corpus de genres différents, dans le but de fournir à la communauté un outil de pré-annotation des chaînes de référence. Nous projetons aussi d'appliquer notre méthode à d'autres langues.

Références

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

BONTCHEVA K., DIMITROV M., MAYNARD D., TABLAN V., CUNNINGHAM H. (2002). Shallow methods for named entity coreference resolution. *Actes de TALN 2002*.

BEAVER D. (2004). The optimization of discourse anaphora. *Linguistics and Philosophy*, 27(1) : 3–56.

CHAROLLES M. (1997). L'encadrement du discours : univers, champs, domaines et espaces. *Cahier de Recherche Linguistique* 6, 1-73.

GEGG-HARRISON W., BYRON D. (2004). PYCOT: An Optimality Theory-based Pronoun Resolution Toolkit. *Actes de LREC 2004*, Lisbonne.

HARTRUMPF S. (2001). Coreference Resolution with Syntactico-Semantic Rules and Corpus Statistics. *Actes de CoNLL (Computational Natural Language Learning Workshop)*.

HERNANDEZ N. (2006). *Description et Détection Automatique de Structures de Textes*. Thèse d'Université, Paris Sud XI.

ION R. (2007). *TTL: A portable framework for tokenization, tagging and lemmatization of large corpora*. Bucharest : Romanian Academy.

KLEIBER G. (1994). *Anaphores et Pronoms.* Louvain-la-Neuve : Duculot.

LONGO L., TODIRASCU A. (2010). Une étude de corpus pour la détection automatique de thèmes. *Actes des 6èmes journées de linguistique de corpus (JLC 09)*, Lorient.

MANUELIAN H. (2003). *Description Définies et Démonstratives : Analyses de Corpus pour la Génération de Textes*. Thèse de doctorat, Nancy 2.

MITKOV R. (2001). Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. *Applied Artificial Intelligence: An International Journal*, 15, 253-276.

SCHNEDECKER C. (2005). Les chaînes de référence dans les portraits journalistiques : éléments de description. *Travaux de Linguistique* 51, 85-133. Duculot.

SALMON-ALT S. (2001). *Référence et Dialogue finalisé : de la linguistique à un modèle opérationnel*. Thèse d'Université, Université H. Poincaré, Nancy.

VICTORRI B. (2005). Le calcul de la référence. *Sémantique et TAL*, Hermès.