ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement

Judith Muzerelle, Anaïs Lefeuvre, Jean-Yves Antoine, Emmanuel Schang, Denis Maurel, Jeanne Villaneau, Iris Eshkol

ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement

Judith Muzerelle[1], Anaïs Lefeuvre[2], Jean-Yves Antoine[2], Emmanuel Schang[1], Denis Maurel[2], Jeanne Villaneau[3], Iris Eshkol[1]

(1) LLL Orléans, Université d'Orléans
(2) Université François Rabelais Tours, LI, 3 place Jean Jaurès, 41000 Blois
(3) IRISA, Université Européenne de Bretagne, 56100 Lorient
Jean-Yves.Antoine@univ-tour.fr, Emmanuel.Schang@univ-orleans.fr, Jeanne.Villaneau@univ-ubs.fr

RÉSUMÉ

Cet article présente la réalisation d'ANCOR, qui constitue par son envergure (453 000 mots) le premier corpus francophone annoté en anaphores et coréférences permettant le développement d'approches centrées sur les données pour la résolution des anaphores et autres traitements de la coréférence. L'annotation a été réalisée sur trois corpus de parole conversationnelle (Accueil_UBS, OTG et ESLO) qui le destinent plus particulièrement au traitement du langage parlé. En l'absence d'équivalent pour le langage écrit, il est toutefois susceptible d'intéresser l'ensemble de la communauté TAL. Par ailleurs, le schéma d'annotation retenu est suffisamment riche pour permettre des études en linguistique de corpus. Le corpus sera diffusé librement à la mi-2013 sous licence Creative Commons BY-NC-SA. Cet article se concentre sur sa mise en œuvre et décrit brièvement quelques résultats obtenus sur la partie déjà annotée de la ressource.

ABSTRACT

ANCOR, the first large French speaking corpus of conversational speech annotated in coreference to be freely available.

This paper presents the first French spoken corpus annotated in coreference whose size (453,000 words) is sufficient to investigate the achievement of data oriented systems of coreference resolution. The annotation was conducted on three different corpora of conversational speech (Accueil_UBS, OTG, ESLO) but this resource can also be interesting for NLP researchers working on written language, considering the lack of a large written French corpus annotated in coreference. We followed a rich annotation scheme which enables also research motivated by linguistic considerations. This corpus will be freely available (Creative Commons BY-NC-SA) around mid-2013. The paper details the achievement of the resource as well as preliminary experiments conducted on the part of the corpus already annotated.

MOTS-CLÉS : Corpus, annotation, coréférence, anaphore, parole conversationnelle

KEYWORDS : Corpus, annotation, coreference, anaphora, conversational speech

1 Introduction

Au cours des deux dernières décennies, le TAL a connu des avancées qui ont conduit à de nombreuses applications dédiées au grand public comme aux professionnels. Parmi celles-ci, la recherche d'information et l'indexation de documents constituent un champ applicatif promis à un bel avenir. La qualité des outils d'indexation ou d'interrogation développés pour ces tâches dépend de leur capacité à résoudre les relations de coréférences présentes dans un (ou plusieurs) texte(s). Un des sauts technologiques auquel est confronté ce domaine est en effet celui du suivi des entités faisant l'objet d'une recherche ou indexation.

L'importance de cette résolution a conduit à l'organisation de multiples campagnes d'évaluation internationales (MUC, ACE, SemEval) ou nationales (DEFT) qui ont accompagné l'évolution des techniques de résolution. Aux travaux initiaux basés sur des méthodes symboliques (Lappin et Leass, 1994) ont succédé des approches plus heuristiques (Mitkov, 1994). Enfin, la mise à disposition de larges corpus annotés en coréférence a ouvert la porte aux approches par apprentissage sur données (Soon et al., 2001, Ng et Cardie, 2002). La plupart des recherches actuelles font ainsi appel à des classifieurs reposant généralement sur un modèle à base de paires (Poesio et al., 2011). Celui-ci consiste à identifier dans un premier temps des paires de mentions coréférentes, et à regrouper ensuite ces paires en clusters de même référence. Il n'existe toutefois pas à l'heure actuelle de corpus francophone libre et d'envergure de taille suffisante pour apprendre un système de résolution efficace. C'est à ce manque que répond le corpus ANCOR. Portant sur le français parlé, il soutient avec ses 418 000 mots la comparaison avec les autres langues de grande diffusion.

2 Etat de l'art

Le corpus que nous présentons a été réalisé par le Laboratoire d'Informatique de l'Université de Tours (LI) et le Laboratoire Ligérien de Linguistique (LLL). Ces deux laboratoires s'intéressent particulièrement à la langue parlée. Il est donc naturel que le corpus porte sur la parole conversationnelle. Le Groupe Aixois de Recherche en Syntaxe a été un des pionniers de l'étude du langage parlé en corpus. Ses travaux fondateurs n'ont malheureusement pas eu pour conséquence le développement ultérieur de ressources informatisées en français. La table 1 présente la liste par idiome des corpus annotés en coréférence de plus 200 kMots (Recasens, 2010). Le français est complètement absent de ce panorama. A notre connaissance, le seul corpus en coréférence disponible en français est le corpus DEDE, centré sur l'étude des descriptions définies. Il ne comporte malheureusement que 48 kMots (Gardent et Manuelian, 2005) et ne peut servir à un apprentissage automatique. De même, le corpus du CRISTAL, de grande envergure, ne peut être considéré par le TAL car il ne code que certaines formes particulières d'anaphore (Tuttin et al. 2000).

Le corpus ANCOR que nous avons constitué vise à répondre à cette situation dans le cas d'un genre linguistique particulier : le français parlé conversationnel. Par sa taille (418 kMots), il ne peut être comparé qu'à deux autres corpus oraux de coréférences d'envergure : Switchboard pour l'anglais américain (200 kMots) et COREA pour le néerlandais. Notons toutefois que seule une partie de cette seconde ressource de 350 kMots

concerne la parole (Hendrickx et al., 2008).

Langue	Corpus	Genre
allemand	TüBa-D/Z	*News* = journaux d'information radio-diffusés (transcription de l'oral)
anglais	ARRAU, Switchboard, ACE, SemEval OntoNotes	*News*, weblog, forum, chat, récit oral, conversation téléphonique
arabe	ACE	*News*
catalan	AnCora-Ca	*News*
chinois (mandarin)	ACE, OntoNotes	*News*
espagnol	ACE, Ancora-Es	*News*
italien	I-CAB	*News*
japonais	NAIST Text	*News*
néerlandais	COREA	*News*, oral, texte encyclopédique
tchèque	PDT	*News*

TABLE 1 – Corpus annotés manuellement en coréférence de plus de 200 000 mots.

3 Elaboration du corpus ANCOR

3.1 Financement et contexte scientifique

La création du corpus ANCOR a été financée en deux étapes. Un premier projet interne au PRES Centre Val-de-Loire a permis de réaliser un premier corpus (C02) de 35 kMots et de valider nos choix d'annotation. Notre souci a été de suivre un schéma d'annotation assez riche pour répondre aux besoins du TAL et des linguistiques. Les premiers résultats obtenus avec CO2 ont ainsi validé l'intérêt de la ressource en linguistique (Schang et al. 2011). Sa taille restait toutefois insuffisante pour constituer une ressource d'apprentissage. Le projet ANCOR, soutenu par la région Centre, nous a précisément permis d'atteindre cet objectif.

3.2 Corpus retenus pour l'annotation

Le corpus ANCOR résulte de l'annotation de trois corpus oraux transcrits sous *Transcriber* (Barras et al., 2001) qui étaient disponibles dans nos laboratoires et diffusés librement :

- ESLO, qui correspond à des entretiens sociolinguistiques (Baude et Dugua 2011, Eshkol-Taravella et al. 2012) ;
- OTG, qui correspond à des dialogues interactifs en présentiel entre des individus et le personnel d'accueil de l'Office du Tourisme de Grenoble ;
- Accueil_UBS, qui correspond à des dialogues interactifs par téléphone recueillis auprès du standard téléphonique d'une université (Nicolas et al., 2002).

Notre objectif a été de représenter une certaine diversité de genres en termes de degré d'interactivité du dialogue. Le corpus ESLO, qui correspond à des entretiens, a une interactivité limitée à la différence des deux autres : le plus souvent, l'enquêteur pose en

effet une question à laquelle s'ensuit un assez long monologue de réponse. La table 2 présente la distribution des corpus de parole dans la ressource annotée.

Corpus Parole	Licence de diffusion	Nb de mots	Durée d'enregistrement
ESLO_ANCOR	Extrait ESLO (CC-BY-NC-SA)	417 kMots	25 heures
OTG	CC-BY-NC-SA	26 kMots	2 heures
Accueil_UBS	CC-BY-NC-SA	10 kMots	1 heure

TABLE 2 – Répartition des corpus oraux annotés dans ANCOR

3.3 Procédure d'annotation

L'annotation a été réalisée sur le logiciel *Glozz* (Mathet et Widlöcher, 2009). Nous n'avons pas retenu *MMAX2* (Müller et Strube, 2006) car son interface a été considérée comme moins conviviale par nos annotateurs. ANCOR sera toutefois diffusé sous les formats *GLOZZ* et *MMAX2,* du fait de la grande diffusion de ce dernier. *Glozz* produit une annotation au format XML reposant sur une DTD que nous avons adaptée à notre schéma d'annotation (cf § 4). Autre intérêt, les annotations sont séparées du corpus source avec lequel elles sont synchronisées. Cette annotation déportée (*stand-off annotation*) permet un enrichissement multi-niveaux du corpus, ce qui est intéressant en termes d'évolutivité. Le principal souci rencontré avec *Glozz* est sa difficulté à gérer de gros fichiers (affichage et gestion mémoire). Si cette limitation n'a pas posé de souci sur OTG et Accueil_UBS (courts dialogues), nous avons dû procéder à un découpage des entretiens ESLO. La forte structuration des entretiens fait toutefois que ce découpage n'a pas détruit de chaînes coréférentielles.

Le corpus ANCOR a fait l'objet d'un codage par deux annotateurs suivi d'une révision, selon une procédure en quatre phases successives :

1) Repérage et caractérisation des entités nommées et autres mentions par un annotateur,

2) Révision croisée du repérage par l'autre annotateur et recherche de consensus,

3) Repérage et caractérisation des relations anaphoriques par un annotateur,

4) Révision finale des relations caractérisées par un superviseur.

Cette démarche séquentielle évite une surcharge cognitive aux codeurs et favorise la cohérence des annotations sur la durée. Annotateurs et superviseurs avaient un bon niveau d'expertise (Master ou doctorat en Sciences du Langage). L'annotation s'est déroulée au rythme de 40 000 mots par homme.mois pour un coût global de construction de 90 000€.

La fiabilité du corpus a été estimée sur une expérience pilote qui a consisté à mesurer l'accord entre 4 experts ayant participé à l'annotation, sur un sondage de 10 fichiers. L'estimation de l'accord inter-annotateur reste une question ouverte dans le cas de la coréférence, du fait des problèmes d'alignement entre annotations (Passoneau, 2004 ; Artstein et Poesio, 2008 ; Matthet et Widlöcher, 2011). Nous proposons de contourner ce problème par le calcul de mesures d'accords successifs sur la délimitation des mentions, l'identification de paires coréférentes et le typage des relations suivant le schéma suivant :

1) Délimitation des mentions : calcul d'accord sur le nombre de mentions retenues,
2) Création d'une annotation en mentions par vote majoritaire pour l'étape suivante,
3) Identification des paires de mentions coréférentes: mesure d'accord sur toutes les paires d'entités suivant une matrice de confusion présence/absence de relation,
4) Création d'une annotation en relations non typées toujours par vote majoritaire
5) Typage des relations de coréférences, et mesure d'accord sur le typage seul.

Cette expérimentation est en cours. Les premiers résultats suggèrent que la fiabilité est acceptable. Nous obtenons pour l'identification des paires une valeur de 0,62 avec les trois métriques κ (Cohen, 1960), π (Scott, 1955) et α (Krippendorff, 2004), valeur très proche du seuil de fiabilité proposé par Cohen. Cette estimation est par ailleurs pénalisée par des problèmes de prévalence et la prise en compte des entités explétives dans l'annotation. Nous réfléchissons à une adaptation de nos mesures pour compenser ces biais et présenter une estimation plus fiable de l'accord inter-annotateur.

4 Schéma d'annotation du corpus ANCOR

Le schéma d'annotation que nous avons proposé cherche de manière classique à identifier pour chaque entité référentielle (ou mention) si elle introduit une nouvelle entité du discours, puis si elle réfère à une entité précédemment mentionnée (coréférence) ou si la référence a une entité précédemment mentionnée dans le texte est nécessaire pour son interprétation (anaphore associative). Il n'existe pas de consensus sur le codage de ces relations. Souvent, l'annotation est adaptée à la tâche étudiée ou à une théorie linguistique sous-jacente. Notre annotation cherche à rester générique et est adaptée aux besoins du TALN en identifiant toutes les entités, isolées ou non. Par contre, nous ne procédons pas à une annotation des propriétés utilisées par les algorithmes de classification. Nous partons du principe que les utilisateurs du corpus pourront procéder à ces prétraitements.

4.1 Repérage des entités référentielles

Nous avons annoté l'ensemble du groupe nominal et pas uniquement sa tête. L'annotation a également concerné les pronoms et les groupes prépositionnels (GP). Dans ce dernier cas, la préposition introductive n'est pas intégrée à l'annotation, mais est prise en compte sous forme d'un attribut associé (GP=YES). Nous avons en outre exclu le pronom *ça* et ses dérivés car il reprend souvent l'ensemble d'un groupe verbal, comme dans l'exemple :

(1) L1 : *Pierre a encore cassé sa voiture.*
L2 : *Venant de lui, ça ne m'étonne pas.*

Ces reprises correspondent à des anaphores abstraites. Comme le notent (Dipper et Zinmeister, 2010), un schéma particulier d'annotation est nécessaire pour décrire ce type de coréférence. Ce type d'annotation dépasse largement le cadre du projet ANCOR.

Nous avons par contre annoté les formes explétives de *il* (cf. *il pleut*). Il est en effet important de repérer ces usages non référentiels qui peuvent tromper les systèmes de résolution. Enfin, dans le cas de structures coordonnées (Mazur et Dale, 2007) ou

enchâssées, nous avons choisi d'identifier le groupe ainsi que chaque membre le composant. Tous ces éléments peuvent en effet ancrer une reprise coréférentielle.

4.2 Délimitation des relations

La délimitation des relations consiste à relier les éléments anaphoriques. Certains travaux privilégient une annotation en chaînes (Gardent et Manuélian, 2005 ; Amsili et al, 2007) c'est-à-dire en « *séquences d'expressions singulières apparaissant dans un contexte telles que si l'une de ces expressions réfère à quelque chose, toutes les autres y réfèrent également* » (Corblin, 2005). Dans le projet ANCOR, il a été décidé de relier toutes les relations à la première mention de l'entité référentielle trouvée dans le texte. Ce choix résulte de tests effectués avec des étudiants de Master Linguistique, qui ont montré un meilleur accord entre annotateurs avec cette approche. Il est en effet apparu que l'annotation en chaîne posait des problèmes délicats pour le dialogue, les annotateurs se trouvant devant des changements de locuteurs pour lesquels la notion de chaîne, pertinente dans la linéarité de l'écrit, devient beaucoup moins évidente à caractériser. Faute de pouvoir inclure (ou exclure) systématiquement d'une chaîne les mentions faites par des locuteurs différents, nos tests ont montré que les annotateurs se trouvaient dans l'impossibilité de trancher de façon nette. Par ailleurs, le codage en première mention rend compte des changements de genre grammatical lors de reprises successives comme dans la séquence "*j'ai une personne qui* (...) *elle téléphone* (...) *c'est un étudiant de L1 ... elle... il...*" où toutes les entités sont coréférentes.

Des arguments d'ordre linguistique ou computationnel peuvent être trouvés en faveur de chaque représentation. C'est pourquoi notre codage sera transformé également en codage en chaîne dans la distribution finale. Notons toutefois que le type de relation (directe, indirecte, pronominale, associative) et l'accord dépendent du choix d'annotation effectué, sans qu'une solution alternative ne nous paraisse envisageable.

4.3 Caractérisation des relations anaphoriques et de leurs entités

L'annotation consiste enfin à décrire par différents traits les entités référentielles et leurs éventuelles relations. Pour les entités nous avons retenu les traits linguistiques suivants :

- **G : Genre** et N : **Nombre**
- **POS : partie du discours** – Ce trait peut prendre les valeurs P (pronom), N (Nom) ou NULL (artefact lié à certaines disfluences)
- **GP : inclusion dans un GP** – Valeur YES (si l'entité est un GP) ou NO (si c'est un GN)
- **EN : entité nommée** – Types retenus dans la campagne d'évaluation ESTER2 (Galliano et al., 2009), à savoir FONC, LOC, PERS, ORG, PROD, TIME, AMOUNT et EVENT. On utilise le type NO si l'entité n'est pas une entité nommée.
- **DEF : définitude** – cet attribut sert à distinguer le caractère défini (DEF), indéfini (INDEF), démonstratif (DEM) ou explétif (EXP) de l'entité.
- **NEW : nouvelle entité du discours** : YES (première mention), NO (entité coréférente avec une autre). Une mention isolée recevra donc toujours le type YES.

Les relations sont caractérisées par un type (trait **TYPE**). Nous distinguons le type *direct (DIR)* dans le cas d'une coréférence entre mentions de même tête nominale (*le bus rouge…. ce grand bus*) ; *indirect (IND)* si les entités coréférentes ont des têtes nominales différentes (*le cabriolet… cette décapotable*) ; *pronominal (PR)* dans le cas particulier de l'anaphore indirecte où la reprise est un pronom (*le cabriolet … il roulait*…) et *associatif (ASSOC)* si les mentions ne sont pas coréférentes mais que l'interprétation de l'une dépend de l'autre (*le village … son clocher*). De même, on retrouve un type associatif pronominal (*ASSOC_PR*).

4.4 Comparaison avec d'autres schémas d'annotation

Notre modèle d'annotation repose est proche de schémas proposés par plusieurs auteurs travaillant sur le langage écrit (van Deemter et Kibble, 2000 ; Vieira et al. 2002). Nous avons en effet le souci que nos travaux puissent être également exploités par des personnes travaillant sur l'écrit. Notre typologie de relations reste relativement simple. Gardent et Manuélian (2005) ont ainsi développé un schéma d'annotation des relations anaphoriques (*bridging*) selon une typologie plus précise. Celle-ci nous semble aller au-delà des besoins actuels du TAL. (Recasens et al., 2011) introduit de son côté la notion de quasi-identité pour des cas décrits comme de la quasi-coréférence et qui sont considérés dans notre schéma comme anaphores associatives. Ces propositions cherchent à réduire les désaccords entre les anaphores associatives et les autres types, mais ne disent rien de ceux entre nouvelle entité du discours et anaphore associative. Cette distinction est pourtant hautement subjective (Vieira et al., 2002). (Ogrodniczuk et al. 2013) ont ainsi rencontré un très faible accord avec un jeu d'annotation intégrant la quasi-identité.

Conclusion

ANCOR est à notre connaissance le premier corpus de français parlé annoté en coréférences diffusé librement et d'envergure suffisante pour permettre un apprentissage automatique. Le LI travaille ainsi au développement d'un système de résolution qui sera appris sur le corpus. Ce système reposera sur BART, une plateforme modulaire et ouverte utilisant le format MMAX comme format d'échange interne (Versley et al. 2008). La richesse d'annotation du corpus permettra également au LLL de conduire des études linguistiques variées sur la coréférence. ANCOR sera diffusé librement sous licence CC BY-NC-SA à la mi-2013. Il sera récupérable sur http://tln.li.univ-tours.fr/Tln_Corpus_Ancor.html.

Références

AMSILI, P., LANDRAGIN, F., ACOSTA, A., BITTAR, A. (2007). Résolution anaphorique : Etat d'une réflexion collective, *Actes Journées d'Etudes de l'ATALA 2007*, pages 1–4.

ARTSTEIN, R., POESIO, M. (2008) Inter-Coder agreement for Computational Linguistics, *Computational Linguistics*, 34, pages 555-596

BARRAS, C., GEOFFROIS, E., WU, Z., LIBERMAN, M. (2001). Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication* 33(1-2), pages 5–22.

BAUDE, O., DUGUA, C. (2011) (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ? *Corpus* ,10, pages 99-118.

ESHKOL-TARAVELLA, I., BAUDE, O., MAUREL, D., HRIBA, L., DUGUA, C., TELLIER, I., (2012) Un grand corpus oral « disponible » : le corpus d'Orléans 1968-2012. *TAL*. 52(3), pages 17-46.

COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, pages 37-46

CORBLIN, F. (2005) Les chaînes de la conversation et les autres. In Gouvard, J.-M. (éd.), *De la langue au style*, Lyon : Presses universitaires de Lyon, pages 233–254.

GARDENT, C. et MANUELIAN, H. (2005). Création d'un corpus annoté de traitement des descriptions définies. *Traitement Automatique des Langues, TAL* , 46(1).

HENDRICKX, I. et al. (2008). A coreference corpus and resolution system for Dutch. *Proc. LREC'2008*.

KRIPPENDORFF., K. (2008). Testing the reliability of content analysis data: what is involved and why. In KRIPPENDORFF, K., ET BLOCH, M.A. (Eds) *The content analysis reader*. Sage Publ..

LAPPIN, S. et LEASS, H.J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*. , 20(4), pages 535-561.

MATHET, Y., WIDLÖCHER, A. (2009). La plate-forme GLOZZ : environnement d'annotation et d'exploration de corpus. *Actes de TALN-2009*, pages 1–10.

MATHET, Y., WIDLÖCHER, A. (2011). Une approche holiste et unifiée de l'alignement et de la mesure d'accord inter-annotateurs. *Actes TALN 2011*, Montpellier, France.

MITKOV, R. (1994). An integrated model for anaphora resolution. *Proc. COLING'94*, Kyoto.

MÜLLER, C., STRUBE, M. (2006). Multi-level annotation of linguistic data with MMAX2. In: Braun, S., Kohn, K., Mukherjee, J., ed., *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Peter Lang, Francfort, Allemagne, pages 197–214.

NG, V. et CARDIE, C. (2002). Improving machine learning approaches to coreference resolution. *Proc. ACL'92*.

NICOLAS, P., LETELLIER-ZARSHENAS, S., SCHADLE, I., ANTOINE, J.-Y., CAELEN, J. (2002). Towards a large corpus of spoken dialogue in French that will be freely available: the "Parole Publique" project and its first realisations. *LREC'2002*. Las Palmas, Espagne. pp. 649-655.

OGRODNICZUK, M., ZAWISŁAWSKA, M., GŁOWINSKA K., SAVARY A. (2013). Interesting Linguistic Features in Coreference Annotation of a Highly Inflectional Language, soumis à *ACL' 2013*.

PASSONEAU, R. (2004) Computing reliability for Co-Reference Annotation. *LREC' 2004*.

PONZETTO, S.P., VERSLEY, Y. (2011) Computational models of anaphora resolution: A survey. Consulté sur : wwwusers.di.uniroma1.it/~ponzetto/pubs/poesio10a.pdf

POESIO, M., PONZETTO, S.P., VERSLEY, Y. (2011) Computational models of anaphora resolution: A survey. Consulté sur : wwwusers.di.uniroma1.it/~ponzetto/pubs/poesio10a.pdf

RECASENS POTAU, M. (2010). Coreference: Theory, Annotation, Resolution and Evaluation. PhD Thesis, Universitat de Barcelona, Catalunya, septembre 2010.

SOON, W.M., NG, H.T. LIM, D.C.Y. (2001). A machine learning approach to coreference resolution in noun phrases. *Computational Linguistics*. , 27(4), pages 521-544.

SCHANG, E., BOYER, A., MUZERELLE, J., ANTOINE, J.-Y., ESHKOL, I., MAUREL, D. (2011). Coreference and anaphoric annotations for spontaneous speech corpus in French, *Proc. Discourse Anaphora and Anaphor Resolution Colloqium, DAARC'2011.*, Faro, Portugal.

SCOTT, W. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinions Quaterly*. 19, pages 321-325.

VAN DEEMTER, K., KIBBLE, R. (2000). On Coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics,* 26(4), pages 629–637.

VERSLEY, Y., PONZETTO, S.P., POESIO, M., EIDELMAN, V., JERN, A., SMITH, J., YANG, X., MOSCHITTI, A. (2008) BART: A Modular Toolkit for Coreference Resolution. *Companion Volume ACL' 08.*

VIEIRA, R., SALMON-ALT, S., SCHANG, E. (2002). Multilingual corpora annotation for processing definite descriptions. *Advances in Natural Language Processing*, pages 721–729.