Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ANCOR

Adèle Désoyer* — **Frédéric Landragin*** — **Isabelle Tellier*** — **Anaïs Lefeuvre**** — **Jean-Yves Antoine****

* *Lattice, CNRS, ENS et Université de Paris 3*
** *LI, Université François Rabelais de Tours*
Adele.Desoyer@gmail.com, Landragin@ens.fr, Isabelle.Tellier@univ-paris3.fr, Anais.Lefeuvre@univ-tours.fr, Jean-Yves.Antoine@univ-tours.fr

RÉSUMÉ. *Cet article présente CROC (Coreference Resolution for Oral Corpus), le premier système de résolution des coréférences en français reposant sur des techniques d'apprentissage automatique. Une des spécificités du système réside dans son apprentissage sur des données exclusivement orales, à savoir ANCOR (anaphore et coréférence dans les corpus oraux), le premier corpus de français oral transcrit annoté en relations anaphoriques. En l'état actuel, le système CROC nécessite un repérage préalable des mentions. Nous détaillons les choix des traits – issus du corpus ou calculés – utilisés par l'apprentissage, et nous présentons un ensemble d'expérimentations avec ces traits. Les scores obtenus sont très proches de ceux de l'état de l'art des systèmes conçus pour l'écrit. Nous concluons alors en donnant des perspectives sur la réalisation d'un système* end-to-end *valable à la fois pour l'oral transcrit et l'écrit.*

ABSTRACT. *We present CROC (Coreference Resolution for Oral Corpus), the first machine learning system for coreference resolution in French. One specific aspect of the system is that it has been trained on data that are exclusively oral, namely ANCOR (ANaphora and Coreference in ORal corpus), the first corpus in oral French with anaphorical relations annotations. In its current state, the CROC system requires pre-annotated mentions. We detail the features that we chose to be used by the learning algorithms, and we present a set of experiments with these features. The scores we obtain are close to those of state-of-the-art systems for written English. Then we give future works on the design of an end-to-end system for oral and written French.*

MOTS-CLÉS : *corpus de dialogues, détection de coréférences, apprentissage, paires de mentions.*

KEYWORDS: *Dialogue corpus, Coreference resolution, Machine learning, Mention-pair model.*

1. Introduction

Depuis les vingt dernières années, la reconnaissance automatique des chaînes de coréférence représente un objet d'étude à part entière du TAL, au cœur de grandes campagnes d'évaluation telles que celles proposées par MUC (*Message Understanding Conference*[1]), ACE (*Automatic Content Extraction*[2]), SemEval (*Semantic Evaluation*[3]) ou CoNLL (*Computationnal Natural Language Learning*[4]). Ces chaînes constituent une unité discursive complexe qui contribue à la cohésion du discours. Les identifier automatiquement oblige à prendre en compte la séquence des phrases qui le composent, et leurs relations. Ce domaine a donné lieu à de nombreux travaux, mais les données sur lesquelles ils se sont fondés (issues des campagnes précédemment citées) étaient jusqu'à présent essentiellement de l'anglais écrit. Les travaux présentés dans cet article ont la particularité de se concentrer sur la reconnaissance automatique de chaînes de coréférence présentes dans de l'oral transcrit français. Cette modalité rend-elle les coréférences plus ou moins fréquentes, explicites et faciles à repérer ? Ce sera tout l'enjeu des expériences que nous avons menées.

Commençons par définir précisément notre objet d'étude. Les chaînes de coréférence s'appuient sur la notion plus restreinte d'anaphore. Cette dernière décrit une procédure référentielle regroupant les phénomènes de renvoi à un antécédent du discours immédiat. Les anaphores sont des relations asymétriques entre un antécédent et une expression anaphorique qui ne peut être interprétée qu'à partir de son antécédent. Dans l'exemple de la figure 1, extrait d'un dialogue du corpus OTG d'ANCOR (que nous présenterons en détail dans cet article), l'anaphorique nominal « le nom » ne peut être interprété qu'à partir de son antécédent « une grande libraire » ; il en va de même pour le pronom « elle ».

— on m'a parlé d'**une grande libraire** mais on se rappelle plus **le nom**
— **Arthaud**
— **Arthaud** peut-être
— **elle** se trouve dans le centre ville

Figure 1. *Mise en évidence d'une chaîne de coréférence dans un dialogue*

Le phénomène plus large de la coréférence se décrit, quant à lui, comme la relation existant entre plusieurs expressions référant à une même entité. Contrairement à l'anaphore qui distinguait strictement ses deux parties, la relation de coréférence est symétrique. Dans l'exemple de la figure 1, l'ensemble des expressions en gras composent une chaîne de coréférence.

1. Voir notamment la tâche sur la coréférence dans MUC-7 en 1998, cf. `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html`.
2. Cf. `http://www.itl.nist.gov/iad/mig//tests/ace/`.
3. Voir notamment la tâche sur la coréférence dans SemEval-2 en 2010, cf. `http://semeval2.fbk.eu/semeval2.php?location=tasks`.
4. Voir notamment la tâche sur la coréférence en 2011 et 2012, cf. `http://conll.cemantix.org/2011/` et `http://conll.cemantix.org/2012/`.

Les systèmes de résolution automatique de la référence sont encore aujourd'hui extrêmement rares s'agissant du français, et même, à notre connaissance, inexistants – à l'exception de systèmes à base de règles tels que celui décrit dans (Trouilleux, 2001) ou RefGen présenté dans (Longo, 2013). L'apprentissage automatique n'avait pu encore être mis en œuvre dans ce contexte, faute jusqu'à présent de corpus annotés et disponibles librement. Le tableau 1 recense les principaux corpus annotés en coréférence et disponibles dans plusieurs langues, d'une taille suffisante pour être exploitables par de tels systèmes. Le français n'y est pas représenté pour les raisons suivantes. Il existe deux corpus s'en approchant. Le premier, un corpus de près d'un million de mots annoté en relations anaphoriques et réalisé conjointement par les laboratoires CRISTAL et XRCE (Tutin *et al.*, 2000), se limite aux pronoms anaphoriques, aux démonstratifs et à quelques ellipses nominales. Il s'avère ainsi bien en deçà de la question de la coréférence nominale. Le second, DEDE (Gardent et Manuélian, 2005), est centré sur l'étude de descriptions définies et ne comporte que 48 000 mots. Cette taille ne permet pas l'exploitation de techniques d'apprentissage automatique, et c'est aussi pourquoi les corpus MUC-6 (Grishman et Sundheim, 1996) et MUC-7 (Hirschman et Chinchor, 1997) n'apparaissent pas non plus dans le tableau 1.

Langue	Corpus	Genre	Taille
Allemand	TüBa-D/Z (Hinrichs *et al.*, 2005)	Informations (News)	800 000
Anglais	ACE-2007	News, dialogues, forums…	300 000
Anglais	OntoNotes (Pradhan *et al.*, 2007)	News, weblogs, dialogues, conversations téléphoniques, flux radio ou télédiffusés	500 000
Chinois	OntoNotes (Pradhan *et al.*, 2007)		400 000
Chinois	ACE-2007	News, weblogs	250 000
Catalan	AnCora-Ca (Recasens, 2010)	Informations	400 000
Espagnol	Ancora-Es (Recasens, 2010)	Informations	400 000
Espagnol	ACE-2007	Informations (News)	200 000
Japonais	NAIST Text (Iida *et al.*, 2007)	Informations	970 000
Hollandais	COREA (Hendrickx *et al.*, 2008)	Informations, parole…	325 000
Tchèque	PDT (Nedoluzhko *et al.*, 2009)	Journaux d'information	800 000
Polonais	PCC (Ogrodniczuk *et al.*, 2013)	Nombreux genres	514 000

Tableau 1. *Principaux corpus annotés en coréférence (taille en nombre de mots)*

Dans cet article, nous présentons le corpus de référence ANCOR (Lefeuvre *et al.*, 2014), « ANaphore et Coréférence dans les Corpus ORaux », et un premier système de résolution des coréférences, CROC, « *Coreference Resolution for Oral Corpus* », entraîné sur ce corpus (Désoyer, 2014). L'objectif de ce travail est donc d'apprendre, à partir d'un corpus de français oral annoté finement en référence et en coréférence, un modèle de résolution automatique de la coréférence. Les résultats nous permettront de dégager des caractéristiques de construction de la référence en français, et plus spécifiquement des phénomènes propres à l'oral.

Dans ce qui suit, nous commençons par présenter les choix qui ont présidé à la constitution du corpus ANCOR, et nous détaillons certaines de ses spécificités. Nous nous attachons ensuite aux systèmes de résolution automatique de la coréférence, au

sein desquels nous situons nos propres modèles. Nos plans d'expérience et nos résultats font l'objet de la dernière partie, précédant la conclusion et les perspectives. Notons que la chaîne de traitement mise en place ne constitue pas un système dit *end-to-end*, puisque nous considérons toujours disposer au départ de l'annotation des unités référentielles. La tâche est donc moins complexe que celle traitée par les systèmes anglo-saxons, mais elle permet de mettre en évidence certains enjeux propres à la phase même de résolution de la coréférence.

2. Contenu du corpus ANCOR

2.1. *Constitution du corpus*

Le corpus ANCOR[5], réalisé dans le cadre du projet du même nom financé par la région Centre, étudie les formes de reprise dans le discours, en annotant les phénomènes anaphoriques et de coréférence sur différents corpus. Le corpus ANCOR ne concerne que la modalité orale. Sans constituer une ressource équilibrée, il ambitionne de représenter une réelle diversité de situations discursives. Il regroupe ainsi l'annotation de quatre corpus de parole spontanée (ESLO_ANCOR, ESLO_CO2, OTG et Accueil_UBS) transcrits sous Transcriber[6]. Deux d'entre eux ont été extraits du corpus ESLO, qui regroupe des entretiens sociolinguistiques présentant un degré d'interactivité faible (Eshkol-Taravella *et al.*, 2012). À l'opposé, les deux autres corpus, OTG et Accueil_UBS (Nicolas *et al.*, 2002), concernent des dialogues homme-homme interactifs. Ces deux derniers corpus diffèrent par le média utilisé : le corpus OTG regroupe des conversations de visu au sein d'un office de tourisme pour OTG, tandis qu'Accueil_UBS a été enregistré dans un standard téléphonique. Au total, le corpus regroupe 488 000 mots et correspond à une durée d'enregistrement de 30,5 heures (tableau 2). Il est donc de taille comparable à ceux présents dans le tableau 1.

Corpus	Situation discursive	Finalisation	Interactivité	Taille et durée
ESLO_ANCOR	Interview	Modérée	Faible	417 kMots – 25 h
ESLO_CO2	Interview	Modérée	Faible	35 kMots – 2,5 h
OTG	Dialogue oral	Très forte	Forte	26 kMots – 2 h
Accueil_UBS	Dialogue téléphonique	Assez forte	Forte	10 kMots – 1 h

Tableau 2. *Contenu du corpus ANCOR. Le degré de finalisation (troisième colonne) correspond à celui de la focalisation du discours sur un but précis.*

Ce corpus a ensuite été annoté manuellement en mentions et relations de coréférence ou anaphoriques à l'aide du logiciel GLOZZ (Mathet et Widlöcher, 2009). On ne détaillera pas ici le schéma d'annotation qui a été adopté (voir (Lefeuvre *et al.*, 2014) pour plus de renseignements). Relevons simplement qu'il a été choisi suffisamment

5. `http://tln.li.univ-tours.fr/Tln_Ancor.html`.
6. `http://trans.sourceforge.net/en/presentation.php`.

fin pour répondre aux études linguistiques en corpus sur la référence, et pour permettre également la mise en œuvre d'approches supervisées pour la résolution automatique des coréférences. Un des objectifs des travaux présentés dans cet article est précisément de voir si une telle annotation riche est suffisante pour obtenir une bonne performance de reconnaissance. Ainsi l'annotation distingue :

1) toutes les mentions nominales ou pronominales du corpus, qu'elles soient ou non référentielles. Ces mentions sont qualifiées par un ensemble de neuf traits :

- type morphosyntaxique, genre, nombre, inclusion ou non dans un groupe prépositionnel, type éventuel d'entité nommée, définitude ;

- caractère spécifique ou générique dans le cas d'une mention référentielle, et enfin attribut binaire précisant si la mention est une nouvelle entité du discours ;

2) toutes les relations anaphoriques ou coréférentielles, identifiées par paires de mentions pointant sur le premier référent du discours correspondant à l'entité coréférentielle. Ces relations sont qualifiées par quatre traits : type de relation (coréférence directe, indirecte ou pronominale, anaphores associatives), accord en genre et en nombre, et enfin identité éventuelle du locuteur entre les deux mentions reliées.

Dans les expériences d'apprentissage, les mentions (avec leurs propriétés listées en 1) seront supposées connues, seules les relations (en 2) devront être identifiées.

Au final, le corpus intègre 115 672 mentions et 51 494 relations (tableau 3) et offre la possibilité de conduire des analyses représentatives même sur des aspects assez rares de la coréférence. La proportion des nouvelles mentions reste stable sur tous les sous-corpus, ce qui suggère que la coréférence est un processus autant guidé par les nécessités de la programmation discursive que par des considérations pragmatiques.

Corpus	ESLO	CO2	OTG	UBS	TOTAL
Mentions (tous types confondus)	97 939	8 399	7 462	1 872	115 672
dont nouvelles mentions (NEW)	*26,8 %*	*32,2 %*	*38,4 %*	*33,7 %*	*28,0 %*
dont mentions coréférentes	*73,2 %*	*67,8 %*	*61,6 %*	*66,3 %*	*72,2 %*
Relations (tous types confondus)	44 597	3 670	2 572	655	51 494

Tableau 3. *Décompte des entités présentes dans ANCOR et ses sous-corpus*

2.2. *Mentions présentes dans le corpus*

Comme nous l'avons précisé, le corpus ANCOR englobe différents sous-corpus correspondant à des contextes interactifs assez variés. Dans un premier temps, il nous a semblé utile de mener une étude variationniste des phénomènes de référence et de coréférence entre ces sous-corpus, afin de voir dans quelle mesure les modèles de résolution qui seront appris sur cette ressource dépendront de sa spécificité. Nos observations distributionnelles sont rassurantes sur le sujet, puisqu'elles dénotent le plus souvent une forte stabilité des résultats d'un corpus à l'autre. Les résultats donnés dans le tableau 4, qui concernent la distribution des mentions référentielles, dénotent éga-

lement une stabilité assez remarquable entre des sous-corpus qui représentent pourtant des genres oraux différents (interview *vs* dialogue oral finalisé). On constate tout d'abord qu'entités nominales et pronominales s'équilibrent toujours fortement. Il semble que la reprise pronominale réponde là encore avant tout à une logique de programmation discursive. Cette observation est à rapprocher des travaux de (Kenny et Huyck, 2011), qui suggèrent que l'usage des pronoms est plus lié à la saillance discursive qu'à la saillance situationnelle. Dans un autre registre, le système de la langue doit être convoqué pour expliquer la quasi-stabilité (entre 25,7 % et 29,9 %) de la proportion de mentions inclues dans un groupe prépositionnel (GP). Cette observation sur la langue générale serait à vérifier en langue de spécialité.

Corpus	ESLO	CO2	OTG	UBS	TOTAL
Entités nominales	48,4 %	51,7 %	52,5 %	51,5 %	48,9 %
Entités pronominales	51,6 %	48,3 %	47,5 %	48,5 %	51,1 %
% de mentions dans un GP	28,0 %	29,9 %	27,8 %	25,7 %	28,1 %
Genre mentions : % masculin	52,8 %	56,7 %	50,5 %	49,9 %	52,9 %
Genre mentions : % féminin	43,9 %	40,2 %	39,3 %	44,4 %	43,3 %
Genre mentions : % inconnu	3,2 %	3,2 %	10,2 %	5,7 %	3,7 %
Nombre mentions : % singulier	65,0 %	68,1 %	66,0 %	83,0 %	65,6 %
Nombre mentions : % pluriel	31,8 %	28,8 %	24,3 %	14,2 %	30,8 %
Nombre mentions : % inconnu	3,2 %	3,2 %	9,6 %	2,8 %	3,6 %
% d'indéfinis	25,1 %	27,3 %	17,2 %	11,9 %	24,5 %
% de définis simples	65,9 %	66,0 %	74,0 %	80,3 %	66,7 %
% de définis démonstratifs	6,9 %	5,2 %	6,2 %	6,5 %	6,7 %
% d'explétifs	2,0 %	1,5 %	2,6 %	1,3 %	2,0 %

Tableau 4. *Étude distributionnelle sur les mentions*

L'accord en genre et en nombre est une caractéristique considérée par l'ensemble des méthodes de résolution de la coréférence (pour une analyse plus précise de cette caractéristique dans le corpus ANCOR, voir (Lefeuvre *et al.*, 2014)). L'analyse du genre des mentions révèle, là encore, une forte stabilité, avec une prédominance légère mais significative du genre masculin. L'existence de mentions de genre inconnu est due à la présence d'entités nommées, de type toponyme ou de sociétés, pour lesquelles la notion de genre n'est pas opérante :

1) La Gacilly est un village charmant du Morbihan.

2) Le Grand Lemps est une petite ville qui peine à maintenir une activité [...].

Le fort taux de mentions de genre inconnu dans le corpus OTG est précisément lié à la prédominance des toponymes dans un corpus recueilli en office de tourisme. Il est difficile de comparer les études portant sur le genre. Sjöblom (2002) observe une prédominance des substantifs masculins (56,5 %) sur les féminins dans l'œuvre de Le Clézio, sans que cette prédominance soit statistiquement significative. À l'opposé, Brunet observe une prédominance du féminin dans l'œuvre de Hugo (53,8 %). Une recherche sur le corpus Frantext donne enfin une prédominance de féminin (56 %) selon Sjöblom.

La situation est plus tranchée du côté du nombre, où le singulier prédomine de manière significative. Là encore, les mentions de nombre inconnu relèvent le plus souvent des toponymes pour qui cette notion n'est généralement pas pertinente. On observe un taux encore plus fort de mentions au singulier dans le corpus Accueil_UBS. Une observation qualitative du corpus semble expliquer cette prédominance du singulier par des dialogues concernant le plus souvent une personne unique que l'on cherche à joindre (corpus standard téléphonique).

Le trait de définitude est également considéré par tous les systèmes de résolution. ANCOR distingue deux types de définis syntaxiques : les définis simples (introduits par l'article défini) et les définis démonstratifs (tableau 4). On observe là encore des régularités notables entre les sous-corpus, ce qui traduit l'absence d'influence sensible du degré d'interactivité sur ce facteur : les définis simples représentent toujours une très forte majorité des mentions. Moins nombreux, les indéfinis restent très fréquents dans tous les corpus. Il est communément admis que les articles définis en français servent à introduire un nom déjà identifié (entité déjà mentionnée dans le discours), facilement identifiable (emplois situationnels) ou bien des types (catégories générales d'êtres ou de choses). Recasens *et al.* (2009) s'interrogent sur les définis qui démarrent une chaîne référentielle et indiquent que les définis en début de chaîne (*chain-starting*) représentent plus de 50 % des cas. Pour l'espagnol, Recasens trouve 73 % de définis en initiale de chaîne, pour le français, Vieira *et al.* (2002) obtiennent 49,6 % des définis classés en nouvelle entité du discours. Une requête sur ANCOR nous permet d'observer que les SN définis introduisent une nouvelle entité du discours dans 53,2 % des cas. Ces observations convergent avec les études citées précédemment.

PERS	LOC	ORG	AMOUNT	TIME	PROD	PROD	PROD
28 856	4 121	1 832	1 649	1 465	1 334	438	201
(72,3 %)	(10,3 %)	(4,6 %)	(4,1 %)	(3,7 %)	(3,4 %)	(1,1 %)	(0,5 %)

Tableau 5. *Distribution des entités nommées par type dans le corpus ANCOR*

Le tableau 5 donne la répartition des entités nommées en fonction de leur type. On observe, comme attendu, une prédominance des personnes (PERS) et des géonymes (LOC). Par son envergure, ANCOR regroupe plus de mille entités nommées, pour la plupart des types. Il s'agit donc d'une ressource potentiellement utile à des travaux spécifiques à la problématique des entités nommées.

Enfin, le tableau 6 présente la répartition des premières mentions des chaînes de coréférence. Sans surprise, l'écrasante majorité des relations anaphoriques ou de coréférence s'ancrent sur une entité nominale : les cataphores, introduites par un pronom, sont très minoritaires.

2.3. *Données quantitatives sur la composition des chaînes*

Nous avons également observé les propriétés des chaînes de coréférence sur différents critères. Les propriétés précédentes étant stables sur l'ensemble des différents

Corpus	ESLO	CO2	OTG	UBS	TOTAL ANCOR
Entité nominale	97,5 %	97,7 %	96,9 %	97,8 %	97,4 %
Pronom (cataphore)	2,5 %	2,3 %	3,1 %	2,2 %	2,6 %

Tableau 6. *Répartition des relations par catégorie de premier référent (ancre)*

	Défini	Indéfini	Démonstratif	Explétif	Total
	Premiers maillons de chaîne				
Nom	57 %	35 %	2 %	0 %	94 %
Pronom	1 %	3 %	2 %	0 %	6 %
	Ensemble de tous les maillons				
Nom	32 %	15 %	1 %	0 %	48 %
Pronom	33 %	11 %	6 %	2 %	52 %

Tableau 7. *Distribution des catégories syntaxiques des maillons*

corpus, nous nous sommes concentrés pour ces mesures sur un sous-ensemble de données représentatives (ensemble de développement décrit en section 4.1, constitué à parts égales des différents sous-corpus). La première observation concerne la proportion considérable de singletons par rapport aux chaînes de deux maillons au minimum : elle atteint 78 % des entités du corpus, contre seulement 22 % de chaînes d'au moins deux maillons. Les trois quarts des entités sont donc isolées, et un système de résolution automatique devra veiller à ne pas les intégrer dans une chaîne. Dans les calculs de distributions présentés en figures 2 et 3, ainsi que dans le tableau 7, portant sur la longueur des chaînes, la distance entre deux maillons et leurs types, seules les chaînes à mentions multiples (à deux maillons au minimum) sont considérées. Les distributions illustrées dans les figures 2 et 3 suivent apparemment toutes les deux une courbe de type Zipf. Le tableau 7 fait, quant à lui, écho aux observations faites en 2.2, puisqu'il démontre que dans une majorité des cas, c'est un défini qui débute les chaînes de coréférence, et que sur l'ensemble des mentions, ce sont les définis et indéfinis qui sont les plus représentés (91 % au total), au détriment des démonstratifs et des explétifs[7].

2.4. *Spécificités des annotations et conséquences pour leur exploitation*

Certains des choix faits pour l'annotation du corpus ANCOR ont des conséquences pour la construction d'un système de reconnaissance de la coréférence par apprentissage automatique. Nous détaillons ici les principaux.

7. Les mentions explétives sont celles qui n'ont aucun ancrage référentiel, comme le pronom *il* dans *il pleut*. Les distributions du tableau nous permettent de vérifier qu'aucun explétif n'a été annoté par erreur en début de chaîne, de même qu'aucune entité nominale n'est explétive.

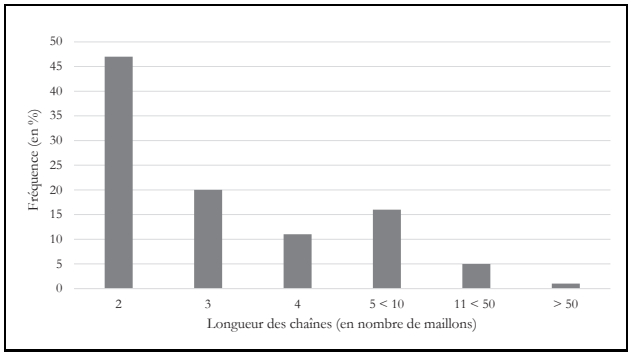

Figure 2. *Distribution des longueurs de chaînes de coréférence*

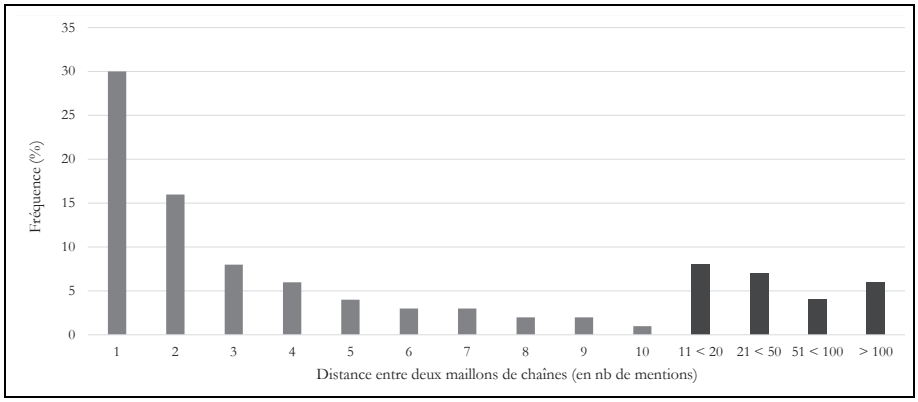

Figure 3. *Distribution des distances séparant deux mentions successives d'une chaîne de coréférence*

Dans ANCOR, le choix a été fait de relier chaque mention d'une entité faisant partie d'une chaîne à la première mention de la chaîne, et non à la précédente. Le fait est à noter car la très grande majorité des systèmes de résolution font le choix de relier chaque maillon d'une chaîne au précédent (ou à leur ensemble) et non au premier. La transformation d'un format en un autre peut se faire aisément, mais au prix de la perte de certaines informations. Par exemple, dans ANCOR, un attribut associé à la relation dit si les deux unités reliées ont la même tête nominale (le type de la relation est alors associé à la valeur *DIRECTE*) : cette information aurait pu constituer un trait pertinent pour l'apprentissage, mais elle n'aura plus aucun sens en changeant de format. À terme, il est prévu que le corpus ANCOR soit distribué aussi bien avec une annotation en première mention (état actuel) qu'en chaîne de coréférence. Une seconde remarque concerne l'annotation des pronoms déictiques dans le corpus. ANCOR étant de l'oral transcrit, leurs occurrences y sont nombreuses, et toutes constituent des unités réfé-

rentielles. En revanche, ces unités ne sont pas intégrées dans une relation quand elles réfèrent à un des interlocuteurs, et sont dans ce cas considérées comme des singletons. Ainsi, les successions de « je » et « tu » dans un dialogue entre deux personnes ne sont pas reliées entre elles, alors qu'elles réfèrent toutes à un même participant. Ce type de coréférence est omniprésent dans le cadre du dialogue oral spontané très interactif. Par son importance quantitative, il aurait pu détourner l'attention des annotateurs vis-à-vis de coréférences plus rares et plus subtiles. Nous avons donc fait le choix de ne pas annoter ce phénomène très balisé qui présente peu d'intérêt linguistique. Dans la perspective d'une application TAL plus couvrante, nous envisageons toutefois de rajouter ces relations à terme.

3. Définition des données d'apprentissage

3.1. *État de l'art des systèmes de résolution de la coréférence*

Les premiers systèmes de résolution automatique de la coréférence traitent la tâche de façon symbolique, avec des règles écrites à la main. Dans les années 1970, la problématique est limitée à la résolution des anaphores pronominales, avec une mise en avant et des pistes pour calculer la saillance (Lappin et Leass, 1994). Elle débouche sur des approches dites *knowledge-poor*, à l'instar de celle de Mitkov (2002) qui, pour seul prétraitement, nécessite une analyse morphosyntaxique et un découpage en *chunks*. À partir de ces données prétraitées, le système commence par repérer les expressions pronominales référentielles, puis, pour chacune d'entre elles, un algorithme combinant une dizaine de règles – parmi lesquelles la saillance, la distance référentielle entre un antécédent et sa reprise ou la répétition lexicale qui privilégie l'antécédent le plus souvent mentionné – attribue à chaque antécédent un score. L'antécédent sélectionné est celui dont le score est le plus élevé. Ces approches ont rapidement trouvé leurs limites et l'essor de la linguistique de corpus a encouragé l'exploitation de données attestées. Les efforts de recherche actuels se concentrent désormais sur des approches fondées sur l'apprentissage supervisé, ce qui nécessite deux prérequis : reformuler l'identification d'une chaîne comme un problème de classification ou d'annotation ; disposer d'un corpus annoté servant à la fois pour l'apprentissage et le test, afin d'évaluer les performances du système ainsi construit.

Différents types d'approches s'opposent quant à la façon de formuler la tâche confiée aux algorithmes d'apprentissage, parmi lesquels :

– les modèles *mention-pair* ou *pairwise* qui sont fondés sur une classification binaire comparant une anaphore à des antécédents potentiels situés dans les phrases précédentes. Concrètement, les exemples fournis au programme sont des paires de mentions (une anaphore et un antécédent potentiel) pour lesquelles l'objectif est de déterminer si elles sont coréférentes ou non. Une anaphore ne pouvant avoir qu'un unique antécédent, une deuxième phase doit déterminer, parmi les paires de mentions classées comme coréférentes, quel est le véritable antécédent d'une anaphore parmi tous ceux qui sont possibles. Différents systèmes ont été implémentés pour cela, parmi

lesquels ceux de Soon *et al.* (2001) et Ng et Cardie (2002), régulièrement utilisés comme système de référence à partir desquels les nouveaux systèmes comparent leur performance. Pour les premiers, l'antécédent sélectionné parmi un ensemble pour une anaphore donnée est celui qui en est le plus proche. Il s'agit d'un regroupement dit *Closest-First* qui, pour chaque anaphore, parcourt l'ensemble du texte vers la gauche, jusqu'à trouver un antécédent ou atteindre le début du texte. Les seconds proposent une alternative à cette approche, dit regroupement *Best-First*, qui sélectionne comme antécédent celui ayant le plus haut score de « probabilité coréférentielle » parmi l'ensemble des précédentes mentions. L'inconvénient de ce type de méthode est la réduction du problème à une série de classifications binaires indépendantes, qui ne prend pas en compte l'ensemble des différents maillons d'une même chaîne de coréférence ;

– les modèles *twin-candidate*, proposés dans (Yang *et al.*, 2003) considèrent également le problème comme une tâche de classification, mais dont les instances sont cette fois composées de trois éléments (*x,* y_i*,* y_j) où *x* est une anaphore et y_i et y_j deux antécédents candidats (y_i étant le plus proche de *x* en termes de distance). L'objectif du modèle est d'établir des critères de comparaison des deux antécédents pour cette anaphore, et de classer l'instance en FIRST si le bon antécédent est y_i et en SECOND si le bon antécédent est y_j. Cette classification alternative est intéressante car elle ne considère plus la résolution de la coréférence comme l'addition de résolutions anaphoriques indépendantes, mais prend en compte l'aspect « concurrentiel » des différents antécédents possibles pour une anaphore ;

– les modèles *mention-ranking* tels que celui décrit dans (Denis, 2007), envisagent non plus d'étiqueter chaque paire de mentions mais de classer l'ensemble des antécédents possibles pour une anaphore donnée selon un processus itératif qui compare successivement cette anaphore à deux antécédents potentiels : à chaque itération, on conserve le meilleur candidat, puis on forme une nouvelle paire de candidats avec ce « gagnant » et un nouveau candidat. L'itération s'arrête lorsqu'il n'y a plus de candidat possible. Une alternative à cette méthode propose de comparer simultanément tous les antécédents possibles pour une anaphore donnée ;

– les modèles *entity-mention* déterminent quant à eux (Yang *et al.*, 2008) la probabilité qu'une expression réfère à une entité ou à une classe d'entités précédemment considérées comme coréferentes (*i.e.* un candidat est comparé à un unique antécédent ou à un cluster contenant toutes les références à une même entité).

Enfin, parmi les travaux les plus récents, certains cherchent à concilier les avantages de chaque méthode, y compris celles à base de règles, en distinguant plusieurs strates de résolution de manière à optimiser les performances en fonction des phénomènes ciblés par chaque strate (Lee *et al.*, 2013).

Nous nous sommes concentrés dans le cadre de cet article sur les stratégies de type *pairwise*, les plus simples à mettre en œuvre tout en permettant d'évaluer assez précisément l'impact de différents attributs sur la qualité de la chaîne reconnue. Il est d'ailleurs souligné dans (Bengtson et Roth, 2008) qu'un simple modèle de classification binaire est capable d'obtenir d'aussi bons résultats qu'un système plus complexe de classement, grâce à la pertinence de son ensemble d'attributs.

3.2. *Résolution de la coréférence comme tâche de classification*

CROC est donc fondé sur une classification binaire opérant sur des paires d'unités référentielles, pour les ranger soit dans la classe des mentions coréférentes (COREF), soit dans celle des mentions non coréférentes (NOT_COREF). La qualité d'un tel système repose sur les données qui lui sont fournies en apprentissage, et particulièrement sur les traits linguistiques (ou *attributs*) décrivant les unités à classer.

	Traits	Valeurs possibles
1	Distance en nombre de phrases entre *i* et *j*	nombre entier
2	*i* est-il un pronom ?	vrai ou faux
3	*j* est-il un pronom ?	vrai ou faux
4	Les chaînes de caractères de *i* et *j* sont-elles égales ?	vrai ou faux
5	*j* est-il un SN défini ?	vrai ou faux
6	*j* est-il un SN démonstratif ?	vrai ou faux
7	*i* et *j* s'accordent-ils en nombre ?	vrai ou faux
8	*i* et *j* s'accordent-ils en genre ?	vrai ou faux
9	*i* et *j* appartiennent-ils à la même classe sémantique ?	vrai ou faux
10	*i* et *j* sont-ils tous deux des noms propres ?	vrai ou faux
11	*i* et *j* sont-ils alias l'un de l'autre ?	vrai ou faux
12	*i* et *j* sont-ils au sein d'une structure appositive ?	vrai ou faux

Tableau 8. *Ensemble des attributs de Soon et al. (2001) pour caractériser un antécédent i et une anaphore j*

Les systèmes de l'état de l'art procédant de la sorte s'inspirent tous de l'ensemble des traits définis dans (Soon *et al.*, 2001), composé de douze attributs présentés dans le tableau 8. Ng et Cardie (2002) ajoutent à ceux-ci quarante et un nouveaux attributs, eux-mêmes repris dans les travaux plus récents. Ces traits se répartissent dans deux familles : les traits non relationnels, qui décrivent une mention d'entité, et les traits relationnels, qui caractérisent la relation unissant les deux mentions d'une paire. Dans ce qui suit, nous décrivons les attributs relevant de ces deux familles que nous avons utilisés, avec la convention suivante : m_1 est considéré comme un antécédent, et m_2 comme une anaphore dont le système doit vérifier qu'elle coréfère ou non à m_1.

3.3. *Traits non relationnels*

Ce type de traits est destiné à décrire chacune des deux mentions d'une paire, indépendamment l'une de l'autre. Les valeurs s'y rapportant apparaîtront donc toujours en double : celles pour m_1, et celles pour m_2.

3.3.1. *Informations morphosyntaxiques*

Les traits 2, 3, 5 et 6 de l'ensemble de (Soon *et al.*, 2001) sont des informations morphosyntaxiques de type non relationnel. Nous les reprenons pour notre propre ensemble, en les adaptant aux possibilités offertes par l'annotation du corpus ANCOR.

3.3.2. *Informations énonciatives*

Le corpus annoté dont nous disposons fournit une information qu'il est difficile de catégoriser, mais qui semble relever de l'énonciation : c'est ce que Denis (2007) décrit comme « le statut discursif de l'expression ». Dans sa thèse, celui-ci développe un système de classification dont la tâche est de déterminer si une mention *m* d'un document introduit ou non une nouvelle entité dans le discours. À l'issue de cette tâche, chaque mention sera annotée de la valeur de sa classe, soit NEW si l'entité référentielle est nouvelle dans le discours, soit OLD si une autre mention l'avait déjà introduite auparavant. Cette information est intégrée au corpus ANCOR sous la forme d'un attribut NEW, qui pour chaque expression peut prendre la valeur YES ou NO.

3.3.3. *Informations sémantiques*

Il s'agit certainement du type d'information parmi les plus compliqués à récupérer automatiquement, puisqu'il nécessite l'intégration de ressources externes telles que des dictionnaires, des thésaurus, des ontologies ou d'autres systèmes de repérage automatique d'entités nommées. L'ensemble de traits des précurseurs Soon *et al.* (2001) intégrait déjà des informations de ce type en associant leurs données à celles de WordNet (Gilbert et Riloff, 2013). Cette ressource qu'est WordNet permet entre autres de répondre au trait 9 du tableau 8, vérifiant que deux mentions d'une paire appartiennent ou non à une même classe sémantique. Nous ne disposons pas de ressources de ce genre pour le français, et aucune annotation de ce type n'apparaît dans le corpus ANCOR. Néanmoins, celui-ci intègre une autre annotation sémantique que des travaux comme ceux de Bengtson et Roth (2008) ou de Recasens (2010) ont introduit dans leur ensemble de traits : celle concernant les entités nommées. L'information nous paraît pertinente car elle pourrait constituer un facteur décisif de classement en négatif lorsque les deux mentions d'une paire ne sont pas de même type : en effet, une expression de type *localisation* ne pourra jamais être coréférente à une expression de type *person*. L'annotation du corpus ANCOR est particulièrement riche concernant le typage des entités nommées, puisqu'il en distingue huit qui constitueront l'ensemble des valeurs possibles des deux nouveaux traits intégrés à notre ensemble[8].

3.4. *Traits relationnels*

Les traits relationnels permettent de comparer les deux mentions d'une paire à partir des précédents traits non relationnels de chacune d'entre elles, ainsi que par différents calculs de distance.

8. L'annotation reprend les types principaux de la campagne d'évaluation ESTER2 (`http://www.afcp-parole.org/camp_eval_systemes_transcription/`), auxquels a été ajouté le type EVENT proposé par le projet QUAERO dans le cadre de la campagne d'évaluation ETAPE (`http://www.afcp-parole.org/etape.html`) qui a fait suite à ESTER2.

3.4.1. *Distances lexicales*

Cette première catégorie s'intéresse à la comparaison des formes linguistiques de m_1 et m_2, à différents degrés, pour créer quatre nouveaux traits. Le premier d'entre eux compare strictement les deux chaînes de caractères, et prend la valeur TRUE uniquement si elles sont exactement identiques. Pour affiner cette première mesure avec des égalités partielles, nous définissons aussi de nouvelles distances lexicales entre les deux chaînes. Un deuxième trait booléen prend ainsi la valeur TRUE si la plus petite des deux mentions est complètement incluse dans la plus grande en tant que chaîne de caractères. Le taux d'inclusion compare, quant à lui, le nombre de tokens partagés au nombre de tokens de la plus petite mention (si les tokens de la plus petite chaîne sont dispersés dans la plus grande, la valeur de ce trait vaudra 1). Le dernier trait s'obtient par le rapport entre l'intersection des deux ensembles de tokens formés par m_1 et m_2 et de leur union. Avec m_1 = « *la grande ville* » et m_2 = « *la ville* », on a ainsi : ID_FORM = FALSE ; ID_SUBFORM = FALSE ; INCL_RATE = 1 et COM_RATE = $\frac{2}{3}$.

3.4.2. *Distances morphosyntaxiques*

Ces traits s'obtiennent en comparant les valeurs des informations morphosyntaxiques de chacune des mentions d'une paire (cf. section 3.3.1). Il s'agit de vérifier la correspondance des genre et nombre, ainsi que celle de la catégorie syntaxique des deux mentions. Les valeurs de ces nouveaux traits sont toutes booléennes, puisque l'égalité est vraie si les valeurs des deux mentions sont identiques, fausse sinon.

3.4.3. *Distances spatiales et syntaxiques*

L'ensemble de référence de Soon *et al.* (2001) contient un attribut comptabilisant le nombre de phrases séparant les deux mentions d'une paire (premier attribut du tableau 8). Les travaux suivants de Recasens et Hovy (2009), Ng et Cardie (2002), Denis (2007), Bengtson et Roth (2008), Recasens (2010) et Stoyanov *et al.* (2010) le reprennent et lui ajoutent un trait du même type, qui calcule cette fois le nombre de paragraphes entre les deux mentions. Ces mêmes travaux, mis à part ceux de Ng et Cardie (2002), intègrent également un trait évaluant la distance entre deux expressions en nombre de mentions ainsi que, pour certains d'entre eux (hormis Denis (2007)), en nombre de mots.

Les données dont nous disposons permettent d'intégrer à notre ensemble d'attributs certaines de ces informations, comme la distance en nombre de mentions et celle en nombre de mots. Il n'est en revanche pas pertinent de calculer celles s'appuyant sur des unités telles que la phrase ou le paragraphe, qui sont propres à l'écrit. Le format d'annotation de Transcriber structure en revanche ses données sous forme de tours de parole, grâce auxquels il est possible de calculer une nouvelle distance spécifique de l'oral. À tous ces attributs, nous ajoutons également un trait décrivant la distance séparant deux mentions en nombre de caractères, pour obtenir un sous-ensemble de quatre traits déterminant divers degrés de distance entre mentions.

Un autre trait, qui relève de la structure syntaxique de surface de l'énoncé, décrit dans les travaux de Recasens et Hovy (2009), Ng et Cardie (2002), Denis (2007), Recasens (2010) et Stoyanov *et al.* (2010) se propose de vérifier si une mention est enchâssée dans une autre plus grande : la mention analysée est-elle une sous-partie d'un syntagme nominal complexe ? Recasens (2010) décrit un trait similaire testant si deux mentions d'une paire sont enchâssées l'une dans l'autre, plus précisément si m_2, la plus à droite des deux selon l'ordre linéaire du texte, compose en partie la précédente m_1. L'annotation d'ANCOR, qui décrit ces inclusions éventuelles dans la portée des mentions, nous permet de définir ce trait qui semble pouvoir jouer un rôle déterminant pour la classification de la paire : en effet si m_2 fait partie de m_1 (*e.g.* « *[la voiture de [mon frère]$_{m_2}$]$_{m_1}$* »), m_2 apporte des informations complémentaires à m_1, mais les deux ne pourront jamais référer à une même entité.

3.4.4. *Distances contextuelles*

Dans cette catégorie, nous insérons le calcul d'attributs vérifiant la correspondance entre les contextes immédiats des deux mentions d'une paire : un premier attribut vérifie l'égalité ou non des tokens à gauche de m_1 et m_2 ; un second fait de même sur les deux tokens à droite. Dans l'exemple « je suis *avec* {**un client**}. Je te rappelle dès que j'aurai terminé *avec* {**lui**} », les deux unités {un client} et {lui} référant à la même entité se trouvent ici au sein du même contexte immédiat : précédé de la préposition *avec*, et suivi d'une ponctuation. Les deux précédents attributs relationnels auront donc pour valeur TRUE.

3.4.5. *Distances énonciatives*

Le premier de ces traits vérifie la correspondance des statuts discursifs (première mention ou déjà introduit dans le discours) des deux mentions.

Un deuxième trait est calculé par comparaison des locuteurs de chacune des deux mentions : l'annotation d'ANCOR intègre ce type d'information, et de nombreux travaux de linguistique de l'oral ont montré l'importance de la coconstruction du dialogue au cours de l'interaction. Cette coconstruction a une influence également sur la coréférence, un locuteur pouvant parfaitement reprendre une mention introduite par son interlocuteur. Ce nouvel attribut, spécifique de nos données orales transcrites, vérifie donc l'identité des locuteurs pour deux mentions d'une paire.

En combinant les traits issus des différents travaux de l'état de l'art avec les annotations disponibles dans le corpus ANCOR, nous avons pu constituer un ensemble de trente traits, décrits dans le tableau 9. La dernière colonne décrit les valeurs possibles pour chacun d'eux : pour les attributs dont la valeur fait partie d'un ensemble restreint de valeurs possibles, l'annotation d'ANCOR ajoute à chaque fois deux valeurs, UNK et NULL. La première est utilisée dans le cas où l'annotateur n'a su se décider sur l'information à renseigner, la seconde dans le cas où l'expression annotée est un artefact lié à certaines disfluences.

	Traits	Définitions	Valeurs possibles
1	m_1_TYPE	catégorie syntaxique de m_1	{N, PR, UNK, NULL}
2	m_2_TYPE	catégorie syntaxique de m_2	{N, PR, UNK, NULL}
3	m_1_DEF	détermination de m_1	{UNK, INDEF, EXPL, DEF_SPLE, DEF_DEM}
4	m_2_DEF	détermination de m_2	
5	m_1_GENRE	genre de m_1	{M, F, UNK, NULL}
6	m_2_GENRE	genre de m_2	{M, F, UNK, NULL}
7	m_1_NOMBRE	nombre de m_1	{SG, PL, UNK, NULL}
8	m_2_NOMBRE	nombre de m_2	{SG, PL, UNK, NULL}
9	m_1_NEW	entité introduite par m_1	{YES, NO, UNK, NULL}
10	m_2_NEW	entité introduite par m_2	{YES, NO, UNK, NULL}
11	m_1_EN	type d'entité de m_1	{PERS, FONC, LOC, ORG, PROD, TIME, NO, AMOUNT, UNK, NULL, EVENT}
12	m_2_EN	type d'entité de m_2	
13	ID_FORM	formes de m_1 et m_2 identiques	{YES, NO, NA}
14	ID_SUBFORM	identité d'une sous-forme	{YES, NO, NA}
15	INCL_RATE	taux d'inclusion des tokens	REAL
16	COM_RATE	taux de tokens communs	REAL
17	ID_DEF	identité de détermination	{YES, NO, NA}
18	ID_TYPE	identité de type	{YES, NO, NA}
19	ID_EN	identité d'entité nommée	{YES, NO, NA}
20	ID_GENRE	identité de genre	{YES, NO, NA}
21	ID_NOMBRE	identité de nombre	{YES, NO, NA}
22	DISTANCE_MENTION	distance en nb. de mentions	REAL
23	DISTANCE_TURN	distance en nb. de tours de parole	REAL
24	DISTANCE_WORD	distance en nb. de mots	REAL
25	DISTANCE_CHAR	distance en nb. de caractères	REAL
26	EMBEDDED	enchâssement de m_2 dans m_1	{YES, NO, NA}
27	ID_PREVIOUS	identité du token précédent	{YES, NO, NA}
28	ID_NEXT	identité du token suivant	{YES, NO, NA}
29	ID_SPK	identité du locuteur	{YES, NO, NA}
30	ID_NEW	identité du statut discrusif	{YES, NO, NA}

Tableau 9. *Ensemble de traits linguistiques*

Soit la chaîne [1,2,3,4,5,6] où chaque chiffre est un maillon					
Annotation des relations dans le corpus (maillon, premier maillon)	[1,2]	[1,3]	[1,4]	[1,5]	[1,6]
Annotation des relations telles que nous les considérons (maillon, maillon précédent)	[1,2]	[2,3]	[3,4]	[4,5]	[5,6]

Figure 4. *Exemple illustrant les différentes possibilités de relier les maillons d'une chaîne de coréférence*

Soit S une suite d'expressions référentielles dans un texte

S = [A_1, b, c, d, A_2, e, A_3]

Soit A la chaîne de coréférence composée des maillons A_1, A_2 et A_3
À partir de l'instance positive [A_1, A_2], sont générées trois instances négatives

[b, A_2], [c, A_2] et [d, A_2]

Figure 5. *Génération d'instances négatives selon Soon et al. (2001).*

4. Expérimentations d'apprentissage et résultats

4.1. *Génération des ensembles de données*

En classification supervisée, la représentation des données est essentielle puisque c'est à partir de celles-ci que les algorithmes construisent leurs modèles. Pour évaluer ensuite les différents systèmes générés, il est nécessaire de bien distinguer les données ayant servi à l'apprentissage de celles destinées au test. À partir de notre corpus, nous distinguons ainsi trois ensembles de données[9] : un ensemble d'apprentissage (constitué d'environ 60 % des données initiales) sur lequel s'appuiera le calcul du modèle ; un ensemble de développement (20 % des données initiales) dont l'observation et la comparaison des résultats pour différents modèles permettront d'optimiser le système ; enfin un ensemble de test (20 % des données initiales) qui évaluera le système final sur de nouvelles données.

La génération de l'ensemble d'apprentissage révèle un premier problème : celui de la distribution des données dans les deux classes des exemples positifs et négatifs. En effet si l'on récupérait toutes les paires de mentions possibles dans un texte, pour former autant d'instances, la classe des non-coréférents serait largement sur-représentée, et biaiserait les évaluations du système ainsi appris. La stratégie proposée dans (Soon *et al.*, 2001), reprise dans (Uryupina, 2004), consiste à récupérer, pour chaque paire de mentions coréférentes [m_1, m_2], toutes les mentions x_i situées entre ces deux maillons, pour construire les paires [x_i, m_2] (cf. illustration figure 5). Néanmoins cette méthode peut, elle aussi, conduire à une surreprésentation des exemples négatifs. Nous avons mené plusieurs expérimentations consistant à garder toutes les paires de maillons internes à une paire coréférente, mais aussi à n'en sélectionner aléatoirement qu'un nombre *n* donné, pour *n* compris entre 1 et 3. Ces premières expériences n'ont cependant pas montré de différences significatives de résultats, en fonction de la différence de proportion entre mentions coréférentes et mentions non coréférentes. Les résultats présentés ici correspondent à des expérimentations réalisées en gardant la même proportion d'environ 20 % de mentions coréférentes et 80 % non coréférentes,

9. Ensembles sélectionnés manuellement en répartissant homogènement des fichiers des différents sous-corpus d'ANCOR dans chacun des trois ensembles.

mais en faisant varier la taille globale du corpus d'apprentissage, pour mesurer son impact. Nous avons donc constitué trois ensembles d'apprentissage distincts avec les caractéristiques suivantes :

– *big_trainingSet* : 142 498 instances dont 24 620 COREF et 117 878 NOT_COREF ;

– *medium_trainingSet* : 101 919 instances dont 17 844 COREF et 84 075 NOT_COREF ;

– *small_trainingSet* : 71 881 instances dont 11 908 COREF et 59 973 NOT_COREF.

Pour ce qui est de la sélection des instances pour les ensembles de développement et de test, la stratégie diffère puisque les chaînes de coréférence n'y sont pas supposées connues, et donc les paires coréférentes non plus. Comme précédemment, sélectionner toutes les paires possibles d'un texte (en récupérant, pour chaque mention, toutes les précédentes pour former une paire) provoquerait une représentation des instances non coréférentes démesurément supérieure à celle des coréférentes. En s'inspirant de la méthode décrite dans (Denis, 2007), dans laquelle les données de test sont construites en récupérant toutes les mentions précédant la courante dans une fenêtre de dix phrases, nous choisissons ici de limiter la génération de paires à une fenêtre de vingt mentions précédentes.

– *developmentSet* : 33 926 instances dont 799 COREF et 33 127 NOT_COREF ;

– *testSet* : 40 497 instances dont 903 COREF et 39 594 NOT_COREF.

4.2. *Plan d'expérimentations*

Différents paramètres entrent en jeu dans la construction du système de détection de chaînes de coréférence ; c'est l'optimisation de la combinaison de ces paramètres qui permettra au système d'améliorer ses performances. Les expérimentations que nous menons ici consistent à générer différents modèles de classification en faisant varier différents facteurs qui relèvent d'une part de la représentation des données, d'autre part de l'algorithme de calcul du modèle. Concernant le premier facteur, la section précédente décrit la mise en place de différents corpus d'apprentissage et de test, pour évaluer l'influence de la taille de ces ensembles sur les résultats finaux. S'ajoute à cela la possibilité de modifier l'ensemble d'attributs initial, pour observer cette fois-ci le niveau d'importance de chacun dans le calcul du modèle. Ainsi, nous distinguons dans nos expériences différents sous-ensembles d'attributs pour évaluer l'apport de certains d'entre eux dans les performances :

– *allFeatureSet* : l'ensemble complet des traits (trente traits au total) ;

– *relationalFeatureSet* : uniquement les traits relationnels (dix-huit traits au total) ;

– *withoutOralFeatureSet* : l'ensemble complet sans les deux traits spécifiques à l'oral, ID_SPK et DISTANCE_TURN (vingt-huit traits au total).

Le second facteur de variation possible des modèles provient de l'algorithme d'apprentissage utilisé. En reprenant les méthodes des systèmes de l'état de l'art, nous

choisissons de tester trois algorithmes distincts : les arbres de décision, les SVM et Naive Bayes tels qu'implémentés dans la plate-forme Weka, avec leurs paramètres par défaut.

Le plan d'expérimentations mis en place consiste à combiner les données des différents corpus d'apprentissage avec les trois ensembles d'attributs produits, puis de fournir chacune de ces représentations de données aux trois algorithmes d'apprentissage. À l'issue de cette classification, les paires sont filtrées pour ne conserver qu'un unique antécédent pour une anaphore. Cette sélection s'appuie sur celle décrite dans les travaux de Soon *et al.* (2001) : il s'agit de la stratégie *Closest-First* qui, lorsqu'une mention a plusieurs antécédents possibles, sélectionne le plus proche à gauche pour former une chaîne. Tous les systèmes ainsi générés seront comparés quantitativement à partir de leurs résultats chiffrés.

4.3. *Évaluations*

La tâche que constitue la résolution de la coréférence est traditionnellement évaluée selon quatre métriques, décrites ici en considérant *T* comme l'ensemble des chaînes avérées et *S* comme l'ensemble des chaînes prédites par le système :

– MUC (dont le nom est issu de la campagne d'évaluation *Message Understanding Conference*) se concentre sur l'évaluation des liens de coréférence qui sont communs à *S* et *T*. Précisément, le rappel correspond au rapport entre le nombre de liens communs à *S* et *T* et le nombre total de liens dans *T*, et la précision au rapport entre le nombre de liens communs aux deux et le nombre total de liens dans *S* : les suppressions sont ainsi pénalisées par le rappel, et les insertions par la précision. Cette prise en compte individuelle de chaque erreur est le principal point faible de cette mesure, qui ne pénalisera pas, par exemple, un système dont l'ensemble final sera composé d'une unique longue chaîne comprenant toutes les mentions : par exemple, deux ensembles *T* = *{{*m_1,m_2,m_3,m_6*}, {*m_4,m_5,m_7*}}* et *S* = *{{*m_1,m_2,m_3,m_4,m_5,m_6,m_7*}}* ne diffèrent que par l'ajout d'un seul lien dans l'ensemble prédit par le système. Le résultat sera donc très bon, alors que pour une interprétation humaine, l'erreur est bien plus importante. S'ajoute à ce problème le fait que la prise en compte des liens comme élément du calcul ne permet pas de considérer les singletons ;

– B^3 : Bagga et Baldwin (1998) considèrent comme unité de base la mention plutôt que le lien. Ainsi pour chaque mention *m* d'un texte, on considère S_m comme la chaîne du système contenant *m* et T_m la chaîne de coréférence contenant *m*. Le rappel est alors le rapport entre le nombre de mentions communes à S_m et T_m et le nombre total de mentions de T_m, la précision à l'inverse est le rapport de l'intersection des deux ensembles et du nombre total de mentions dans S_m. Les résultats proposés par B^3 sont significativement plus représentatifs de la qualité du système évalué, d'une part parce que les singletons sont pris en compte (et ils constituent dans notre cas la plus grande partie des données de test), d'autre part parce qu'un système qui renverrait une unique chaîne finale serait cette fois sévèrement sanctionné ;

– CEAF, développé dans les travaux de Luo (2005), est fondée sur l'entité, c'est-à-dire la référence commune à tous les maillons d'une chaîne de coréférence (le nom complet de la mesure est d'ailleurs *Constrained Entity Aligned F-Measure*). La méthode de calcul est cette fois plus complexe, et surtout bien plus longue, puisqu'il s'agit de trouver, pour chaque chaîne de *S*, la chaîne de *T* la plus proche. Cette sélection s'opère par mesure de similarité entre les deux chaînes (*i.e.* le nombre de mentions communes), sachant que dès qu'une chaîne de *S* a été associée à une chaîne de *T*, elle ne peut plus être associée à aucune autre chaîne (une chaîne du système correspond donc à une unique chaîne de référence). Une première partie de l'algorithme de calcul recherche donc les meilleures correspondances possibles entre chaînes prédites et chaînes de coréférence, puis une seconde phase détermine, à partir de ces correspondances, quelle est la meilleure correspondance globale de l'ensemble de toutes les chaînes ;

– BLANC (pour *BiLateral Assessment of Noun-phrase Coreference*) est la métrique la plus récente mise au point dans les travaux de Recasens (2010), dont la vocation est de considérer conjointement les liens de coréférence et de non-coréférence. Concrètement, deux paramètres sont pris en compte avec d'une part, les réponses du système pour une paire, et d'autre part, les réponses de référence : ces deux paramètres ayant chacun deux valeurs possibles (positif ou négatif pour le premier, vrai ou faux pour le second), l'ensemble forme une matrice de confusion sur laquelle s'appuie le calcul de BLANC, dont la formule de F-mesure est la moyenne des F-mesures des deux classes. La pertinence de cette mesure tient à la prise en compte de la surreprésentation des instances négatives par rapport aux positives dans un corpus, et met en évidence la capacité du système évalué à caractériser ces instances négatives comme telles.

À titre de repères, notons que les meilleurs systèmes conçus pour la langue anglaise en *end-to-end* (c'est-à-dire sans connaître *a priori* les positions des mentions dans le corpus de test) atteignent des scores de 70 à 80 pour la métrique MUC, de 70 à 80 pour B^3, de 65 à 75 pour CEAF, et de 70 à 75 pour BLANC (Désoyer, 2014).

Plusieurs séries de tests nous permettent d'évaluer les différents systèmes mis en place en phase d'expérimentation, afin de déterminer celui présentant les meilleures performances :

– **série 1** : les trois algorithmes sont appris sur l'ensemble *small_trainingSet*, pour chacun des trois ensembles d'attributs ;

– **série 2** : SVM et C4.5 sont appris sur l'ensemble *medium_trainingSet*, pour chacun des trois ensembles d'attributs ;

– **série 3** : SVM et C4.5 sont appris sur l'ensemble *big_trainingSet*, pour chacun des trois ensembles d'attributs.

Le tableau 10 présente les résultats obtenus sur l'ensemble de développement selon toutes les mesures, pour les différents modèles et pour les trois séries d'expériences. À l'issue de ces expérimentations, le système présentant les meilleurs résultats est celui utilisé pour détecter les chaînes de coréférence de l'ensemble de test. Les valeurs obtenues dans ce dernier cas sont cohérentes avec les précédentes.

Nous remarquons tout d'abord qu'avec le plus petit des corpus d'entraînement, Naive Bayes se comporte significativement moins bien que les deux autres algorithmes d'apprentissage. Cela justifie que nous l'ayons écarté pour les expériences menées sur les corpus plus volumineux. L'impact de la quantité de données d'entraînement n'est pas convaincant : pour SVM et C4.5, les résultats sur *small_trainingSet* sont souvent les meilleurs. Au-delà, il est possible que l'effet bien connu du surapprentissage (au-delà d'une certaine quantité de données, les modèles sont surspécialisés sur ces données et généralisent mal sur des données nouvelles) se fasse déjà sentir. Il est possible aussi que notre choix d'apprendre en mélangeant les quatre types initiaux de sous-corpus pénalise les modèles : des modèles spécifiques de chacun de ces sous-corpus, plus homogènes, se comporteraient peut-être mieux.

La prise en compte de l'ensemble de tous les traits (*all_FeatureSet*) semble dans tous les cas bénéfique. Les seuls traits relationnels aboutissent en effet dans la plupart des cas à l'apprentissage de modèles moins performants. Le fait de ne pas prendre en compte les deux traits spécifiques de l'oral est apparemment assez peu pénalisant : les différences de performances semblent minimes et pas toujours orientées de la même façon. Si l'on fait la moyenne des quatre métriques, les résultats avec ou sans ces attributs sont quasiment identiques. Il est possible, ici aussi, que ces résultats seraient à nuancer si nous avions distingué les différents sous-corpus, qui varient quant à leur degré d'interactivité. La distance en tours de parole entre deux mentions successives est donc sans doute différente d'un sous-corpus à un autre, et utiliser ce trait en mélangeant les sous-corpus restreint son impact. Nous observons d'ailleurs, dans les arbres de décision obtenus, que ces traits spécifiques ne figurent pas dans les plus hauts niveaux (c'est-à-dire parmi les attributs les plus discriminants). Ceux figurant à cette place sont plutôt : la distance entre les mentions (ou, aussi, les distances en nombres de mots ou de caractères), l'identité de genre, le caractère d'entités nommées des mentions, le caractère emboîté ou non des mentions.

Enfin, pour ce qui est des algorithmes : sans surprise, c'est SVM qui obtient en général les meilleurs résultats. Les performances atteintes sont dans l'ordre de grandeur des systèmes de l'état de l'art de l'anglais, sauf que CROC suppose connues les mentions initiales dans le corpus de test, alors que ce n'est pas le cas des systèmes *end-to-end*. Le modèle finalement intégré au système de résolution final est celui calculé par SVM sur l'ensemble complet d'attributs et le corpus d'apprentissage *small_trainingSet*, puisque la moyenne de ses quatre métriques est la plus haute (71,9 %). Les résultats de ce système sur l'ensemble de test sont présentés dans le tableau 11.

De manière générale, il est extrêmement délicat de comparer nos résultats à ceux déjà connus, tant les données varient : la langue des corpus, la modalité (oral *vs* écrit), le codage des coréférences (le fait de ne pas considérer les pronoms ou les entités désignant les interlocuteurs comme faisant partie d'une chaîne, par exemple), les traits disponibles, etc., sont différents. CROC se veut plutôt une base nouvelle, à laquelle devront se comparer les futurs autres systèmes de reconnaissance de la coréférence dédiés au français.

		small_trainingSet			*medium_trSet*		*big_trSet*	
		NBayes	SVM	C4.5	SVM	C4.5	SVM	C4.5
all-FeatureSet	MUC	52,95	58,60	59,92	59,48	**64,59**	59,23	63,33
	B^3	51,81	**84,20**	78,39	82,83	77,60	82,50	77,90
	CEAF	42,41	**78,02**	71,36	76,83	70,92	76,30	71,34
	BLANC	51,68	**66,93**	65,14	66,46	66,90	66,13	67,32
relational-FeatureSet	MUC	47,51	39,71	**50,69**	38,42	48,21	39,03	48,22
	B^3	34,32	76,25	**78,41**	75,55	32,18	75,87	32,31
	CEAF	29,91	63,57	**69,61**	62,76	27,83	62,22	27,86
	BLANC	47,40	61,43	**65,61**	60,71	23,06	60,10	24,55
notOral-FeatureSet	MUC	52,46	58,83	59,29	59,09	61,31	59,64	**63,37**
	B^3	55,16	**83,95**	79,38	82,63	77,95	82,35	78,62
	CEAF	44,83	**77,74**	72,39	76,57	71,28	76,03	72,04
	BLANC	54,60	66,44	65,47	66,19	**76,00**	65,69	67,00

Tableau 10. *Résultats des modèles appris avec* small_trainingSet*,* medium_trainingSet *et* big_trainingSet *respectivement sur l'ensemble de développement*

testSet			
MUC	B^3	CEAF	BLANC
63,45	83,76	79,14	67,43

Tableau 11. *Résultats du système final sur l'ensemble de test*

5. Conclusion et perspectives

Depuis que la tâche de résolution de la coréférence occupe une place importante dans les problématiques de traitement automatique des langues, beaucoup de travaux se sont attachés à développer des systèmes de détection de chaînes de coréférence pour l'anglais. Très peu cependant, mis à part celui décrit dans (Longo, 2013), ont étudié le phénomène et ses pistes de résolution automatique sur le français, de fait ce travail présente l'un de ces premiers systèmes appris automatiquement sur un corpus annoté. En plus de cette évolution de langue, c'est une distinction de canal qui caractérise ce travail, puisque à ce jour aucun modèle de résolution fondé sur l'apprentissage supervisé n'avait été développé spécifiquement pour l'oral transcrit.

Les résultats d'évaluation obtenus par notre modèle de résolution, bien qu'assez proches de ceux de certains systèmes de l'état de l'art, sont peu généralisables en envisageant de tester le système sur des données non annotées en mentions. Néanmoins, ces résultats expérimentaux nous ont permis d'observer certaines propriétés du phénomène étudié ainsi que d'envisager certaines pistes de travail pour améliorer les performances du modèle de classification, étendre ses capacités à celles d'un système

end-to-end, et pour en compléter l'évaluation. Pour obtenir un tel système *end-to-end* en français, il faudrait coupler CROC avec un étiqueteur POS, un reconnaisseur d'entités nommées, et divers autres outils capables d'identifier les genres et nombres des mentions, par exemple.

Il est difficile en l'état actuel de nos expériences de mesurer l'impact spécifique des différents traits utilisés. Une perspective de ce travail serait de s'attacher précisément à cette sélection d'attributs, par exemple *via* une méthode de sélection ascendante qui évaluerait un modèle appris sur un ensemble ne contenant qu'un trait, puis ajouterait de manière incrémentale un nouveau trait à l'ensemble, en ne le conservant que si les résultats de classification sont meilleurs que pour l'ensemble précédent.

Dans ce travail, la tâche de classification ne permet de ranger les instances que sous deux classes : soit COREF, soit NOT_COREF. En procédant ainsi, nous ne distinguons pas les différentes formes de reprises telles qu'annotées dans le corpus ANCOR, et ne prenons pas en compte le fait que chacune d'entre elles est susceptible d'avoir des propriétés qui lui sont propres. Notamment, pour l'anaphore pronominale, on pourrait supposer que la distance entre les deux mentions d'une paire ne doit pas excéder un certain seuil. Plusieurs méthodes sont envisageables pour intégrer cette considération : d'une part, modifier les paramètres de classification pour que les différentes classes correspondent aux différents types de reprises, à savoir une anaphore fidèle (lexicalement), infidèle, pronominale et associative, d'autre part, s'inspirer des recherches de Denis (2007) qui propose d'apprendre des modèles spécifiques pour chaque type de reprises, à partir des mêmes traits qui se verront attribuer des poids différents en fonction du type d'expressions pris en compte.

L'impact de la spécificité des sous-corpus (le degré variable d'interactivité, par exemple) n'a pas non plus pour l'instant pu être pris en compte, puisque les différents sous-corpus ont été mélangés dans nos différents ensembles (d'apprentissage et de test). D'autres séries d'expériences pourraient être menées sur les sous-corpus initiaux. Leur taille encore un peu limitée semblait un obstacle, mais nos expériences montrent que des corpus d'apprentissage réduits peuvent néanmoins permettre d'atteindre de bonnes performances.

En tout état de cause, ANCOR constitue une nouvelle référence pour l'étude des chaînes de coréférence en français, et CROC est désormais sa baseline associée.

Remerciements

Le corpus ANCOR, ou plus exactement ANCOR_Centre, a été constitué dans le cadre du projet du même nom, « ANaphore et Coréférence dans les Corpus ORaux », projet soutenu par la région Centre. Nous remercions l'ensemble des participants à ce projet. Le système CROC a été conçu dans le cadre du projet ANR ORFEO, « Outils et recherches sur le français écrit et oral », dont nous remercions également les participants. Nous tenons enfin à remercier les relecteurs anonymes pour leurs commentaires pertinents et enrichissants.

6. Bibliographie

Bagga A., Baldwin B., « Entity-based Cross-document Coreferencing Using the Vector Space Model », *Proceedings of ACL'98*, p. 79-85, 1998.

Bengtson E., Roth D., « Understanding the Value of Features for Coreference Resolution », *Proceedings of EMNLP 2010*, p. 236-243, 2008.

Denis P., New Learning Models for Robust Reference Resolution, PhD thesis, University of Texas at Austin, 2007.

Désoyer A.*, « Apprentissage d'un modèle de résolution automatique de la coréférence à partir d'un corpus de français oral »*, Master's thesis, Université Paris Ouest, 2014.

Eshkol-Taravella I., Baude O., Maurel D., Hriba L., Dugua C., Tellier I., « Un grand corpus oral « disponible » : le corpus d'Orléans 1968–2012 », *TAL*, vol. 52, n° 3, p. 17-46, 2012.

Gardent C., Manuélian H., « Création d'un corpus annoté pour le traitement des descriptions définies », *TAL*, vol. 46, n° 1, p. 115-139, 2005.

Gilbert N., Riloff E., « Domain-Specific Coreference Resolution with Lexicalized Features », *Proceedings of ACL 2013*, p. 81-86, 2013.

Grishman R., Sundheim B., « Message understanding conference - 6 : A brief history », *Proceedings of the International Conference on Computational Linguistics*, 1996.

Hendrickx I., Bouma G., Coppens F., Daelemans W., Hoste V., Kloosterman G., Mineur A.-M., Vloet J. V. D., Verschelde J.-L., « A Coreference Corpus and Resolution System for Dutch », *Proceedings of LREC 2008*, 2008.

Hinrichs E., Kübler S., Naumann K., Zinsmeister H., « Recent developments in linguistic annotations of the TüBa-D/Z Treebank », *Proceedings of the 27th Meeting of the German Linguistic Association*, 2005.

Hirschman L., Chinchor N., « MUC-7 Coreference Task Definition », *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1997.

Iida R., Komachi M., Inui K., Matsumoto Y., « Annotating a Japanese text corpus with predicate-argument and coreference relations », *Proceedings of Linguistic Annotation Workshop*, p. 132-139, 2007.

Kenny I., Huyck C., « Resolution of Anaphoric and Exophoric Definite Referring Expressions in a Situated Language Environment », *Proceedings of DAARC 2011*, 2011.

Lappin S., Leass H. J., « An algorithm for pronominal anaphora resolution », *Computational Linguistics*, vol. 20, p. 535-561, 1994.

Lee H., Chang A., Peirsman Y., Chambers N., Surdeanu M., Jurafsky D., « Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules », *Computational Linguistics*, vol. 39, n° 4, p. 885-916, 2013.

Lefeuvre A., Antoine J.-Y., Schang E., « Le corpus ANCOR_Centre et son outil de requêtage : application à l'étude de l'accord en genre et en nombre dans les coréférences et anaphores en français parlé », *Actes du quatrième congrès mondial de linguistique française*, 2014.

Longo L., Vers des moteurs de recherche intelligents : un outil de détection automatique de thèmes, PhD thesis, Université de Strasbourg, 2013.

Luo X., « On Coreference Resolution Performance Metrics », *Proceedings of Human Language Technology - Empirical Methods in Natural Language Processing (EMNLP 2005)*, 2005.

Mathet Y., Widlöcher A., « La plate-forme GLOZZ : Environnement d'annotation et d'exploration de corpus », *Actes de TALN*, ATALA, p. 1-10, 2009.

Mitkov R., *Anaphora Resolution*, Pearson Education, 2002.

Nedoluzhko A., Mírovský J., Ocelák R., Pergler J., « Extended Coreferential Relations and Bridging Anaphora in the Prague Dependency Treebank », *Proceedings of DAARC 2009*, AU-KBC Research Centre, Anna University, Chennai, Goa, India, p. 1-16, 2009.

Ng V., Cardie C., « Improving Machine Learning Approcahes to Corefrence Resolution », *Proceedings of ACL'02*, p. 104-111, 2002.

Nicolas P., Letellier-Zarshenas S., Schadle I., Antoine J.-Y., Caelen J., « Towards a large corpus of spoken dialogue in French that will be freely available : the "Parole Publique" project and its first realisations », *Proceedings of LREC 2002*, 2002.

Ogrodniczuk M., Głowińska K., Kopeć M., Savary A., Zawisławska M., « Polish coreference corpus », *Proceedings of the 6th Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics*, p. 494-498, 2013.

Pradhan S., Ramshaw L., Weischedel R., MacBride J., Micciula L., « Unrestricted coreference : identifying entities and events in OntoNotes », *Proc. of ICSC'07*, p. 446-453, 2007.

Recasens M., Coreference : Theory, Resolution, Annotation and Evaluation, PhD thesis, University of Barcelona, 2010.

Recasens M., Hovy E., « A Deeper Look into Features in Coreference Resolution », *Proceedings of DAARC 2009*, p. 29-42, 2009.

Recasens M., Martí M., Taulé M., « First-mention Definites : More than Exceptional Cases », *The Fruits of Empirical Linguistics*, De Gruyter, 2009.

Sjöblom M., L'écriture de J.M.G. Le Clezio, PhD thesis, Université de Nice, 2002.

Soon W. M., Ng H. T., Lim D. C. Y., « A Machine Learning Approach to Coreference Resolution of Noun Phrases », *Computational Linguistics*, vol. 27, n° 4, p. 521-544, 2001.

Stoyanov V., Cardie C., Gilbert N., Riloff E., Buttler D., Hysom D., Reconcile : A Coreference Resolution Research Platform, Technical report, 2010.

Trouilleux F., Identification des reprises et interprétation automatique des expressions pronominales dans des textes en français, PhD thesis, Université Blaise Pascal, 2001.

Tutin A., Trouilleux F., Clouzot C., Gaussier E., Zaenen A., Rayot S., Antoniadis G., « Annotating a large corpus with anaphoric links », *Proceedings of DAARC 2000*, 2000.

Uryupina O., « Linguistically Motivated Sample Selection for Coreference Resolution », *Proceedings of DAARC 2004*, 2004.

Vieira R., Salmon-Alt S., Schang E., « Multilingual Corpora Annotation for Processing Definite Descriptions », *Proceedings of Portugal for Natural language Processing*, 2002.

Yang X., Su J., Lang J., Tan C. L., Liu T., Li S., « An Entity-Mention Model for Coreference Resolution with Inductive Logic Programming », *Proc. of ACL'08*, p. 843-851, 2008.

Yang X., Zhou G., Su J., Tan C. L., « Coreference Resolution Using Competition Learning Approach », *Proceedings of ACL'03*, p. 176-183, 2003.