TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Vers la détection des dislocations à gauche dans les transcriptions automatiques du Français parlé

Corinna Anderson[1] Christophe Cerisara[1] Claire Gardent[1]
(1) LORIA-CNRS UMR 7503, Campus Scientifique, Vandoeuvre-les-Nancy
{andersoc,cerisara,gardent}@loria.fr

Résumé. Ce travail prend place dans le cadre plus général du développement d'une plate-forme d'analyse syntaxique du français parlé. Nous décrivons la conception d'un modèle automatique pour résoudre le lien anaphorique présent dans les dislocations à gauche dans un corpus de français parlé radiophonique. La détection de ces structures devrait permettre à terme d'améliorer notre analyseur syntaxique en enrichissant les informations prises en compte dans nos modèles automatiques. La résolution du lien anaphorique est réalisée en deux étapes : un premier niveau à base de règles filtre les configurations candidates, et un second niveau s'appuie sur un modèle appris selon le critère du maximum d'entropie. Une évaluation expérimentale réalisée par validation croisée sur un corpus annoté manuellement donne une F-mesure de l'ordre de 40%.

Abstract. Left dislocations are an important distinguishing feature of spoken French. In this paper, we present a hybrid approach for detecting the coreferential link that holds between left-dislocated elements and the coreferential pronoun occurring further on in the sentence. The approach combines a symbolic graph rewrite step with a maximum entropy classifier and achieves around 40% F-score. We conjecture that developing such approaches could contribute to the general anaphora resolution task and help improve parsers trained on corpora enriched with left dislocation anaphoric links.

Mots-clés : Détection des dislocations à gauche, Maximum Entropy, français parlé.

Keywords: Left dislocation detection, Maximum Entropy, spoken French.

1 Introduction

Les dislocations sont considérées depuis longtemps comme des caractéristiques importantes du Français parlé (De Cat, 2007; Delais-Roussarie *et al.*, 2004; Hirschbuhler, 1975; Lambrecht, 1994). Bien qu'elles apparaissent plus fréquemment dans la parole spontanée, il n'est pas rare de les trouver également dans de nombreux autres registres du français oral, et ce depuis au moins le XVII^{ème} siècle (De Cat, 2007; Blanche-Benveniste, 1997).

Une caractéristique liée à la définition des dislocations à gauche est qu'elles impliquent un lien coréférentiel entre l'élément disloqué and un pronom situé plus loin dans la phrase. Ainsi, le traitement des dislocations à gauche contribue au domaine plus général qui est celui de la résolution des anaphores. De plus, l'annotation automatique ou semi-automatique de ces liens coréférentiels constitue une information potentiellement utile pour améliorer la qualité des analyseurs syntaxiques automatiques de l'oral.

Nous nous intéressons dans cet article à ce problème et présentons une approche hybride de résolution des liens de coréférence entre les éléments disloqués à gauche et le pronom correspondant. Nous décrivons au paragraphe 2 la syntaxe des dislocations à gauche en l'illustrant avec des exemples extraits du corpus radiophonique ESTER du français parlé, initialement conçu pour les campagnes d'évaluation nationales des systèmes de transcription automatique de la parole (Gravier *et al.*, 2004). Notre méthodologie est ensuite décrite au paragraphe 3, ainsi que les résultats expérimentaux. Le paragraphe 4 conclut l'article en présentant notamment quelques perspectives.

2 Dislocation à gauche : syntaxe et coréférence

Une dislocation à gauche peut être définie comme une expression située dans un voisinage gauche d'une proposition qui contient un pronom présentant un lien de coréférence avec cette expression, comme l'illustre l'exemple (Ex.1) dans le cas d'une phrase simple, et les exemples (Ex.2,3) pour une proposition subordonnée. Notons qu'une séparation hiérarchique entre le constituant disloqué et le pronom comme dans (Ex.3) n'a aucun effet sur la coréférence.

Ex.1 : [ma question] **elle** est simplement par rapport au respect des uns et des autres
Ex.2 : je crois que [l' état d' esprit que ça dénote] **c'** est pousse toi de là que je m' y mette quoi
Ex.3 : et tous les jours je constate que [le véhicule qui me suit] quelques fois j' ai bien du mal à lire [**sa** plaque d' immatriculation] car je ne le voit pas il est trop près

La dislocation à gauche présente souvent un contour intonatif caractéristique avec une proéminence prosodique de l'élément disloqué, éventuellement perçue comme étant suivie d'un court silence (De Cat, 2007). Toutefois, nous ne nous traiterons dans le reste de cet article que des aspects syntaxiques des dislocations.

Syntaxe Nous noterons dans la suite respectivement l'élément disloqué à gauche et l'expression interne à la proposition par les acronymes EDG et EIP. En français, l'élément disloqué à gauche, bien qu'il soit typiquement nominal, peut appartenir à différentes catégories syntaxiques : phrase/syntagme simple (Ex.4,6) ou complexe (Ex.5,7), pronom demonstratif ou "emphatique" (Ex.8,9), syntagme prépositionel (Ex.10), proposition infinitive (Ex.11), adjectif, voire un mot d'une toute autre catégorie comme dans (Ex.12).

Ex.4 : oh [cette vague] effectivement **elle** existe
Ex.5 : mais [ce qui est euh très dangereux] **c'** est de ne pas avoir son attention euh sur la route
Ex.6 : tout n' est pas rose entre guillemets euh [l' éclatement du Front National] je m' **en** réjouis
Ex.7 : et pour elle [le seul partenaire euh qui existe qui soit digne de considération que se soit en terme d' ennemi ou en terme plus positif] **ce** sont les Etats-Unis
Ex.8 : et [ça] par contre on ne **le** verbalise pas
Ex.9 : donc euh [moi] **j'** ai vraiment appris à avoir confiance en moi à travers euh le théâtre quoi
Ex.10 : eh ben [à France-Inter] en effet vous **y** êtes
Ex.11 : [payer ces mêmes impôts] **c'** est déjà possible sur le net
Ex.12 : [universel] **c'** est notre mot du jour

L'EDG peut être lié avec un (Ex.14) ou plusieurs (Ex.13) pronoms parmi ceux apparaissant dans une même phrase.

Ex.13 : je sens que [cette décision] je dois vous **la** donner rapidement puisque vous **l'**attendez depuis longtemps
Ex.14 : mais [la situation sociale] on le voit bien **elle** est très tendue avec un appel à la grève générale

L'élément EIP est en général un pronom personnel ou démonstratif, mais ce rôle peut également être tenu par un épithète, nom commun (Ex.15), ou article possessif (Ex.3).

Ex.15 : [Droits de l'Homme] le numéro un chinois aborde précisément **le sujet** dans une interview

Dans nos expérimentations, nous nous focalisons sur les dislocations à gauche dont l'élément disloqué est nominal, pronominal ou infinitif. De plus, l'élément EIP doit être un pronom personnel, possessif ou démonstratif [1], ce qui représente 92 % des instances de dislocations à gauche dans notre corpus. Les 8 % de dislocations à gauche restantes, comme (Ex.8,9), seront comptabilisées comme des erreurs systématiques pour nos approches.

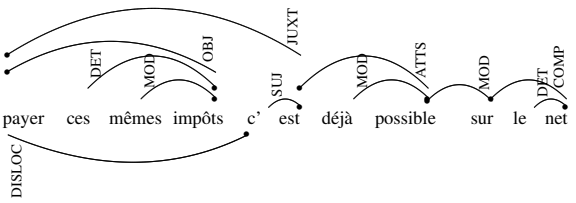

FIG. 1 – Exemple d'annotation d'une dislocation à gauche : les dépendances au-dessus de la phrase correspondent aux dépendances syntaxiques obtenues automatiquement ; l'annotation de la dislocation à gauche apparaît sous la phrase. Le petit rond noir à l'extrémité des arcs représentent la tête, ou gouverneur, de la relation.

Coréférence Dans une phrase avec dislocation, l'EDG et l'EIP partagent une seule référence, mais avec des fonctions differentes : lorsque l'EDG exprime l'identité référentielle en l'annonçant comme topic, l'EIP marque le rôle syntaxique dans la phrase (Lambrecht, 1994). L'EIP peut avoir un rôle quelconque ; on a vu des exemples d'objet direct (Ex.3,8,13), indirect (Ex.6) et modifieur (Ex.10). Néanmoins, 89 % des dislocations à gauche dans

[1] De même, nous ne considérons pas non plus les inversions de sujet complexes comme des instances de dislocations à gauche, car à la différence des véritables dislocations, il n'existe pas de phrase équivalente contenant l'expression nominale complexe sans le pronom.

notre corpus correspondent au sujet d'une proposition principale ou subordonnée ; des chiffres similaires (82.5 % à 85,5 %) ont été découverts par Antoine et Goulian (2001) pour tout type d'antéposition, dislocation et extraction à gauche dans les dialogues, sans distinguer les dislocations des autres structures.

À la différence de la plupart des relations de coréférence dans la phrase, il y a peu de contraintes quant au rôle ou la position à laquelle le pronom doit apparaître. Les deux éléments EDG et EIP peuvent être séparés par une distance arbitraire, à la fois linéairement et structurellement (c'est-à-dire qu'ils peuvent apparaître à tous les niveaux d'une hiérarchie de la phrase). Ainsi, il n'existe pas d'indices structurels simples, comme par exemple le fait d'être argument ou dépendant d'un certain verbe, permettant d'identifier les pronoms participant à la dislocation à gauche. En fait, le pronom référentiel d'une dislocation se comporte exactement comme d'autres pronoms coréférents d'expressions introduites précédemment dans le discours ou dans un contexte extralinguistique ; leur domaine est la phrase entière.

3 Méthodologie et validation expérimentale

Notre objectif est de résoudre automatiquement les liens de coréférence entre les éléments EDG et EIP à partir des seules transcriptions de la parole en français, annotées par un analyseur syntaxique automatique. Nous procédons en deux étapes. Tout d'abord, un filtre symbolique est appliqué pour extraire automatiquement un ensemble de candidats possibles pour les dislocations à gauche à partir du corpus initial. Ensuite, un classifieur basé sur le principe de maximum d'entropie, qui ne requiert pas d'hypothèse d'indépendance entre les indices en entrée, est utilisé pour sélectionner, parmi ces candidats, ceux correspondant le plus vraisemblablement à des dislocations à gauche. Il y a donc deux classes possibles, selon qu'il s'agisse d'une dislocation à gauche ou non.

3.1 Corpus et annotations

Corpus de travail Le corpus choisi pour entraîner et tester notre système est extrait du corpus ESTER, qui contient des enregistrements de radios francophones et est utilisé dans les campagnes d'évaluation nationale des systèmes de transcription automatique (Gravier *et al.*, 2004). Une analyse réalisée sur un échantillon de 10 000 mots de ce corpus montre qu'il contient environ 60 % de parole "préparée", c'est-à-dire prononcée par un journaliste, et 40 % de parole plus spontanée, correspondant à des interviews et commentaires de personnes invitées ou enregistrées par téléphone. Les dislocations à gauche sont en règle général considérées comme caractéristiques de la parole spontanée, mais nous observons que 4 ‰ des mots de notre corpus font partie d'une dislocation, quel que soit le type de parole.

Ponctuation, annotations syntaxiques et coréférentielles Dans une grande partie du corpus ESTER original, les transcriptions sont manuellement annotées avec de la ponctuation. Nous utilisons donc les trois symboles de ponctuation [. ? !] caractéristiques des fins de phrase pour segmenter automatiquement le corpus en phrases. Ensuite, nous supprimons tous les symboles de ponctuation du texte, afin de se rapprocher le plus possible des conditions obtenues en sortie d'un système de transcription automatique, qui ne produit aucun symbole de ponctuation. Nous utilisons alors le Treetagger (Schmid, 1994) pour annoter automatiquement les classes morphosyntaxiques des mots et le MaltParser (Nivre *et al.*, 2007) pour ajouter les arbres de dépendances. Le MaltParser a préalablement été entraîné sur 50 000 mots du corpus ESTER annotés manuellement en dépendances syntaxiques (Cerisara *et al.*, 2010). Les performances de cet analyseur atteignent un LAS (Labeled Attachment Score) de 76 % sur un corpus de test également extrait de ESTER. Les phrases ainsi préparées sont alors annotées manuellement en dislocations, en ajoutant un lien de coréférence de la tête du syntagme disloqué vers le pronom référentiel ; ainsi, les mots *payer* et *c'* sont liés (Fig.1), alors que l'élément disloqué est en fait le syntagme entier *payer ces memes impots*. Cette tâche d'annotation est réalisée grâce au logiciel JSafran (Cerisara *et al.*, 2010), qui gère des graphes d'annotation sur plusieurs niveaux. Les dislocations à gauche et à droite sont toutes deux ainsi annotées, mais seules les dislocations à gauche sont utilisées dans la suite de ce travail.

Le corpus final obtenu est composé de 1735 phrases contenant 35 000 mots. Il contient 84 dislocations à gauche dans 79 phrases. Toutes les expériences suivantes sont réalisées par validation croisée en 10 parties, qui est une configuration standard fréquemment utilisée lorsque la taille du corpus ne permet pas d'extraire deux corpus indépendants pour l'apprentissage et le test.

3.2 Détection des liens coréférentiels des dislocations à gauche

Filtre symbolique Nous avons conçu des règles de réécriture de graphes pour identifier les candidats EDG-EIP possibles et ajouter pour chaque candidat un lien coréférentiel de EDG vers EIP. Ces règles s'appuient sur des expressions régulières sur les graphes exploitant les mots, leur position, leur classe morphosyntaxique et leurs dépendances syntaxiques. Intuitivement, ces règles comparent les classes morphosyntaxiques des mots avec une liste prédéfinie pour ne retenir que les éléments EDG potentiels, puis cherchent des pronoms situés plus loin dans la phrase. Les classes retenues sont : les noms propres et communs, les verbes à l'infinitif, les pronoms démonstratifs ("ça") et possessifs ("le sien"). Notons que les classes morphosyntaxiques sont obtenues automatiquement, et certaines d'entre elles sont donc erronées, ce qui constitue une source d'erreur pour notre système de détection et pour l'analyseur syntaxique. Une autre condition concerne l'existence d'un sujet après l'élément EDG, sujet qui peut précéder ou coïncider avec le pronom EIP. Ce sujet permet de situer la frontière gauche d'une proposition finie, et donc du domaine du pronom candidat de la dislocation qui le précède. Cette approche est motivée par plusieurs observations. En premier lieu, toutes les phrases en français ont un sujet explicite qui marque la frontière gauche de la proposition. De plus, les sujets sont relativement bien identifiés par l'analyseur syntaxique, ce qui limite l'impact de cette source d'erreurs. Les sujets potentiels multiples, notamment avec les dislocations à gauche et les clitiques compléments d'objet ambigus, ainsi que les modifieurs contenant un nom, sont une source de sur-identification de candidats de dislocation, mais aussi l'un des indicateurs importants pour l'identification des EDG. Ce premier filtre identifie environ 20 000 candidats potentiels qui sont transmis à la seconde étape.

Classification automatique des candidats disloqués à gauche Nous avons utilisé la librairie OpenNLP [2] pour entraîner un modèle à maximum d'entropie sur le corpus décrit précédemment. Seuls les candidats retenus par le premier filtre sont considérés dans ce modèle, afin d'exploiter au mieux les paramètres du modèle en n'apprenant que les exemples ambigus et pertinents. De fait, des expériences préliminaires ont montré que, sans ce premier filtre, les performances du modèle chutent de près de 10 % absolus. Le rôle du premier filtre est donc important, car il permet au classifieur de se focaliser sur les zones importantes de l'espace des données. Les observations utilisées dans le modèle à maximum d'entropie sont les suivantes : (1) Les classes morphosyntaxiques de EDG et EIP ; (2) Les étiquettes des dépendances partant de EDG de EIP ; (3) La distance linéaire entre EDG et EIP ; et (4) La position relative du gouverneur syntaxique de EDG.

Les observations (1) sont triviales. Les indices (2) exploitent la structure arborescente en dépendances syntaxiques. La distance linéaire entre EDG et EIP est exprimée en nombre de mots : elle représente intuitivement le fait que les dislocations proches, comme celles impliquant des sujets et objets clitiques, sont plus fréquentes que les dislocations lointaines. Afin de limiter le nombre de paramètres du modèle, ces distances sont échantillonnées et peuvent prendre seulement 4 valeurs : 1, 2, 3 ou 4+. L'indice (4) est calculé en comparant la position du gouverneur GOV_{EDG} relativement à EDG et EIP, et peut prendre 4 valeurs :

$$GOV_{EDG} < EDG \qquad EDG < GOV_{EDG} < EIP \qquad GOV_{EDG} = EIP \qquad GOV_{EDG} > EIP$$

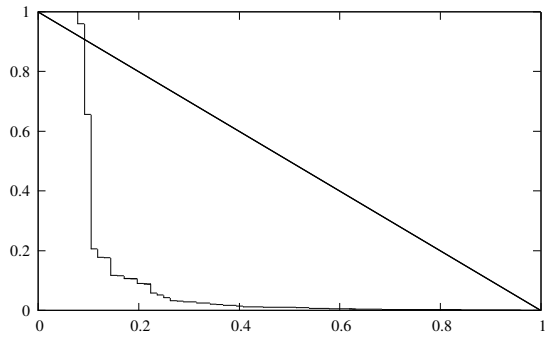

FIG. 2 – Courbe DET (Detection Error Tradeoff) de la détection des dislocations à gauche. L'axe des X représente le taux de faux rejets et l'axe des Y celui des fausses acceptations.

[2] http://incubator.apache.org/opennlp

3.3 Résultats expérimentaux

Les métriques d'évaluation choisies sont (1) le rappel-précision classique, (2) la courbe DET (*Detection Error Tradeoff* (Martin *et al.*, 1997)), qui donne un aperçu global des performances pour tous les seuils possibles, et (3) l'erreur égale (EER) qui représente le point de la courbe DET pour lequel spécificité et sensibilité sont égales.

Performance globales La figure 2 affiche la courbe DET de détection des dislocations à gauche à la sortie du classifieur. Cette courbe est d'autant meilleure qu'elle se rapproche des axes. La partie de la courbe en bas à droite correspond à des seuils de détection très sélectifs qui ne retiennent que très peu de dislocations, mais seulement les plus vraisemblables. La partie de la courbe en haut à gauche correspond à des seuils très permissifs qui acceptent de nombreux candidats, mais qui n'en omettent idéalement aucun. En pratique, la partie gauche de la courbe est translatée de 8 % vers la droite, ce qui correspond aux erreurs (faux négatifs) systématiques du premier filtre.

Une valeur de rappel et précision peut être calculée en chaque point de la courbe. Nous donnons les trois valeurs les plus intéressantes dans le tableau suivant :

Position sur la figure 2	Précision	Rappel	F-mesure
Près de l'extrémité gauche	1 %	92 %	2 %
Au centre	43 %	43 %	43 %
Près de l'extrémité droite	50 %	10 %	17 %

La F-mesure obtenue en choisissant au hasard pour chaque candidat si c'est une dislocation ou non, selon la probabilité a priori de chaque cas, est de 2 %. D'autre part, le faible nombre de dislocations crée un intervalle de confiance statistique assez large : la F-mesure au centre de la courbe est ainsi comprise entre 32 % et 53 %.

Utilité des observations Le tableau suivant présente les performances du détecteur en taux d'erreurs égales selon les types des observations.

Observations	Equal Error Rate
Classes morphosyntaxiques de EDG et EIP	27.7%
+ étiquettes des dépendances	23.2%
+ distance entre EDG et EIP	17.6%
+ position relative du gouverneur de EDG	13.5%

TAB. 1 – Performances de détection selon le type des observations.

Cette expérience est utile pour valider les différentes catégories d'indices proposés, et montrer qu'aucun d'entre eux n'est inutile. Elle permet également de montrer que chaque type d'indices apporte une information différente des autres, au moins en partie. Ces résultats confirment ainsi expérimentalement l'intuition selon laquelle les dislocations à gauche affectent et/ou dépendent de manière significative d'au moins trois niveaux différents : la morphosyntaxe, les structures syntaxiques et l'ordre et la position des mots dans la phrase.

Discussion La détection des dislocations à gauche est relativement similaire à la problématique de résolution des anaphores, notamment parce qu'il n'y a pas de relation syntaxique directe entre les éléments coréférents. C'est pourquoi l'architecture du système proposé dans ce travail suit les principes généraux des systèmes les plus récents conçus pour la résolution des pronoms, tels que décrits dans (Poesio *et al.*, 2010). Toutefois, il existe également des différences fondamentales entre la tâche de détection des dislocations à gauche et les autres systèmes de résolution d'anaphores. En particulier :

– Alors que la résolution d'anaphores vise à lier ensemble toutes les instances de la même entité, nous nous focalisons seulement sur une résolution locale au sein d'une même phrase en liant l'élément disloqué au pronom référentiel associé dans la phrase.
– Tout comme dans le problème général de résolution des pronoms, nous voulons également discriminer les différents pronoms qui font référence au même antécédent des autres pronoms qui, soit ne possèdent aucun référent, comme dans "il y a du pain", ou alors possèdent un antécédent implicite. Mais nous voulons de plus discriminer les pronoms qui sont impliqués dans une dislocation de ceux qui ne le sont pas.

– (Poesio *et al.*, 2010) identifie trois types d'indices utiles pour la résolution d'anaphore : la syntaxe, des connaissances de "sens commun" et la saillance cognitive. Dans ce travail, nous élaborons des indices similaires en nous contraignant à exploiter uniquement ceux qui peuvent être produits à la sortie d'un système de reconnaissance automatique de la parole en français, et en particulier issus d'une analyse syntaxique.

4 Conclusions

Nous avons présenté dans cet article la conception et l'implémentation d'un système de résolution des liens de coréférence pour les cas de dislocations à gauche dans un corpus de français parlé. Ce système détecte les dislocations à gauche en deux étapes. La première étape utilise des règles de transformation de graphes définies de manière experte pour filtrer les éléments candidats de la phrase pouvant potentiellement constituer une dislocation à gauche. Ce filtre utilise essentiellement des contraintes sur les classes morphosyntaxiques des candidats et leurs positions dans l'énoncé decoupé / la phrase découpée. La deuxième étape utilise un modèle basé sur le principe du maximum d'entropie pour classifier chacun de ces candidats comme relevant effectivement d'une dislocation ou non. Ce modèle exploite seulement des indices syntaxiques et lexicaux. Les expériences réalisées sur une sous-partie du corpus radiophonique français ESTER montrent que notre système atteint une F-mesure de 43 %.

Ce système peut être amélioré de nombreuses façons. Tout d'abord, d'autres indices classiquement utilisés pour la résolution de coréférence peuvent être ajoutés, notamment la prosodie ainsi que les préférences lexicales des arguments des prédicats. Augmenter la taille du corpus d'apprentissage devrait également améliorer les performances du modèle, notamment en y incluant plus d'exemples répresentatifs et plusieurs sous-types connus de dislocation qui ne figurent pas dans notre échantillon de corpus. Une fois que des performances satisfaisantes auront ainsi été atteintes, nous envisagerons alors d'intégrer les informations produites par le système de détection des dislocations à gauche en tant que nouveaux indices dans notre modèle d'analyse syntaxique, ce qui devrait avoir un impact positif sur les performances d'analyse syntaxique du français parlé spontané. Le processus résultant, de l'analyse vers la détection des dislocations, puis à nouveau vers l'analyse pourrait être itéré jusqu'à convergence. Finalement, nous pourrons enfin tester ce système intégré sur de véritables transcriptions automatiques de la parole. Mais ces objectifs à long terme requièrent de mener auparavant à bien des recherches visant à améliorer la robustesse de l'analyse syntaxique aux erreurs de reconnaissance de la parole.

Références

BLANCHE-BENVENISTE (1997). *Approches de la langue parlée en français*. Paris : Ophry.

CERISARA C., GARDENT C. & ANDERSON C. (2010). Building and exploiting a dependency treebank for French radio broadcasts. In *Proc. Intl Workshop on Treebanks and Linguistic Theories (TLT)*, Tartu, Estonia.

DE CAT C. (2007). *French Dislocation. Interpretation, Syntax, Acquisition*. Oxford :OUP.

DELAIS-ROUSSARIE E., DOETJES J. & SLEEMAN P. (2004). *Handbook of French Semantics*, chapter Dislocation, p. 501–529. Stanford : CLSI Publications.

GRAVIER G., BONASTRE J.-F., GALLIANO S., GEOFFROIS E., TAIT K. M. & CHOUKRI K. (2004). ESTER, une campagne d'évaluation des systèmes d'indexation d'émissions radiophoniques. In *Proc. JEP*, Fez.

HIRSCHBUHLER P. (1975). On the source of lefthand NPs in French. *Linguistic Inquiry*, p. 155–165.

LAMBRECHT K. (1994). *Information Structure and Sentence Form : Topic, Focus and the Mental Representations of Discourse Referents*. Cambridge University Press.

MARTIN A., DODDINGTON G., KAMM T., ORDOWSKI M. & PRZYBOCKI M. (1997). The DET curve in assessment of detection task performance. In *Proc. Eurospeech*, p. 1895–1898.

NIVRE J., HALL J., NILSSON J., CHANEV A., ERYIGIT G., KÜBLER S., MARINOV S. & MARSI E. (2007). MaltParser : a language-independent system for data-driven dependency parsing. *Natural Language Engineering*, **13**(2), 95–135.

POESIO M., PONZETTO S. P. & VERSLEY Y. (2010). Computational models of anaphora resolution : A survey. to be published ; available at `http://clic.cimec.unitn.it/massimo/Publications/lilt.pdf`.

SCHMID H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proc. International Conference on New Methods in Language Processing*, p. 44–49.