Un modèle simple du coût cognitif de la résolution des anaphores

Olga Seminck[1, 2]
(1) Laboratoire de Linguistique Formelle, 8 place Paul Ricoeur, 75013 Paris, France
(2) Laboratoire d'Informatique de Paris Nord, 99 avenue Jean-Baptiste Clément, 93430 Villetaneuse, France
olga.seminck@cri-paris.org

RÉSUMÉ

Nous présentons un travail en cours sur un projet de recherche en TAL et en psycholinguistique. Le but de notre projet est de modéliser le coût cognitif que représente la résolution d'anaphores. Nous voulons obtenir une mesure du coût cognitif continue et incrémentale qui peut, à un stade de recherche plus avancé, être corrélée avec des mesures d'occulométrie sur corpus. Pour cela, nous proposons une modélisation inspirée par des techniques venues du TAL. Nous utilisons un solveur d'anaphores probabiliste basé sur l'algorithme *couples de mentions* et la notion d'entropie pour établir une mesure du coût cognitif des anaphores. Ensuite, nous montrons par des visualisations quelles sont les prédictions de cette première modélisation pour les pronoms personnels de troisième personne dans le corpus ANCOR Centre.

ABSTRACT

A Simple Model of Cognitive Cost of Anaphora Resolution

We propose work in progress on an interdisciplinary project between NLP and psycholinguistics. The goal of our project is to model the cognitive cost of anaphora resolution. We want to obtain an incremental and continuous measure of cognitive processing cost that can, in a more advanced stage of the project, be correlated with eye-tracking data on corpus. We propose a model that is inspired by NLP techniques. We use a probabilistic mention-pair anaphora resolver and the notion of entropy to establish a measure of cognitive cost for anaphora. Then, we visualize the predictions of this first modelisation for third person personal pronouns in the ANCOR Centre corpus.

MOTS-CLÉS : Modèle Probabiliste, Psycholinguistique, Entropie, Apprentissage Supervisé, Coréférence, Pronoms.

KEYWORDS: Probabilistic Model, Psycholinguistics, Entropy, Supervised Learning, Coreference, Pronouns.

1 Introduction

Lorsque les humains résolvent une anaphore, et ils sont amenés à le faire fréquemment, car c'est un phénomène massif en langue, ils font appel à de nombreuses connaissances et utilisent des heuristiques variées. Ceci a été mis en évidence par de nombreux travaux en psycholinguistique où l'influence d'une multitude de facteurs a été étudiée, comme par exemple la causalité implicite, les rôles grammaticaux des candidats antécédents ou l'influence des connecteurs de discours (Caramazza *et al.*, 1977; Colonna *et al.*, 2015; Kehler & Rohde, 2013). Le coût cognitif joue un rôle important dans l'étude de ces facteurs. En effet, en mesurant le coût cognitif en fonction des types d'anaphore

ou des contextes linguistiques, on peut étudier comment les humains résolvent les anaphores. Le coût cognitif est mesuré indirectement en utilisant des paradigmes expérimentaux qui enregistrent des temps de réaction (par exemple Caramazza *et al.*, 1977) ou en utilisant des enregistrements occulométriques (par exemple Koornneef, 2008), voire en ayant recours à des procédés d'imagerie cérébrale (par exemple Nieuwland & Van Berkum, 2008). On dispose donc à l'heure actuelle d'un ensemble de données sur le coût cognitif de cette tâche.

L'objectif global du projet dont nous présentons ici des résultats préliminaires, est d'élaborer un modèle computationnel cognitivement motivé de résolution anaphorique qui pourra simuler les performances des sujets humains en termes de coût cognitif. L'idée est qu'en étudiant les caractéristiques des modèles computationnels qui simulent le mieux le fonctionnement des sujets humains (en termes de facteurs pris en compte, information nécessaire à l'apprentissage, algorithmes de résolution, etc.), on pourra d'un côté apporter à la psycholinguistique de nouveaux moyens pour départager ou d'élaborer des théories, et d'une autre côté bénéficier d'une nouvelle source d'inspiration pour les applications en TAL.

Notre projet s'inspire de travaux portant sur l'analyse syntaxique automatique qui visaient le même type d'objectif (Hale, 2001, 2003, 2006; Levy, 2008). Plus précisément, il a été développé des modèles computationnels de l'analyse syntaxique qui ont la propriété d'être incrémentaux et probabilistes, ce qui les permet, à chaque étape de l'analyse, c'est-à-dire pour chaque mot d'une phrase, de mesurer un coût. Le coût, basé sur les probabilités du modèle d'analyse syntaxique, a pu être corrélé avec des temps de lectures sur corpus (par exemple Demberg & Keller, 2008; Frank, 2013), montrant ainsi sa pertinence vis-à-vis le comportement humain. Nous voulons élaborer un modèle de coût pour les anaphores ayant ces mêmes propriétés : être incrémental et probabiliste, pour pouvoir à terme le corréler aux temps de lectures.

Ce que nous proposons dans cet article est une première tentative de modélisation de ce coût des anaphores. Pour ce faire, nous prenons comme point de départ les modèles de coût cognitifs proposés pour la syntaxe. Nous décrivons notamment la théorie de la surprise (Hale, 2001) et l'hypothèse de la réduction d'entropie (Hale, 2003, 2006) qui sont à la base de notre propre démarche. Ensuite, nous présentons un système de résolution d'anaphores venant du TAL, fonctionnant par *couples de mentions* [1] probabiliste (Soon *et al.*, 2001), dont nous utilisons les probabilités pour calculer un coût cognitif de l'anaphore basée sur l'entropie (Shannon & Weaver, 1949). Comme notre recherche est toujours en cours, nous n'avons pas encore évalué les prédictions du modèle vis-à-vis du coût cognitif mesuré pour les humains. Cependant, nous proposons une visualisation de notre mesure qui nous permet de faire une première évaluation qualitative et nous présentons des perspectives qu'ouvre ce premier travail.

2 Modélisation computationnelle du coût cognitif

Dans cette section, nous présentons les travaux qui sont à la base de notre démarche : la théorie de la surprise (Hale, 2001) et l'hypothèse de la réduction d'entropie (Hale, 2003, 2006). Après avoir étudié ces deux théories, nous présenterons des travaux qui proposent d'élaborer les mesures de coût cognitif de la syntaxe, pour couvrir plus de phénomènes linguistiques et obtenir une meilleure corrélation avec des temps de lecture sur corpus.

1. Nous utilisons ce terme pour traduire les noms anglais de cet algorithme : *pairwise* et *mention-pair*.

FIGURE 1 – Extrait d'une grammaire hors contexte imaginaire

	Règle			Probabilité
1.	S	→	NP VP	1,0
2.	NP	→	Det N	0,7
3.	NP	→	ProperName	0,3

2.1 La théorie de la surprise

La *surprise* est une propriété d'un événement qui a une probabilité (Sheldon *et al.*, 1998). Intuitivement, si nous faisons une expérience et que nous obtenons un événement E, nous sommes surpris d'obtenir ce résultat si $p(E)$ est toute petite. Par contre, si $p(E)$ est très grande, nous ne sommes pas vraiment surpris. La surprise est donc une fonction qui est inversement liée à la probabilité d'un événement (voir l'équation (1)). Ainsi, un événement d'une grande probabilité aura une surprise moins élevée qu'un événement d'une faible probabilité.

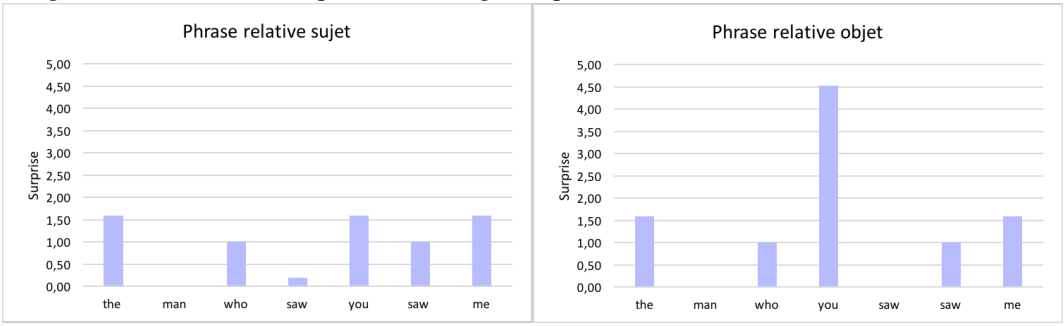
$$Surprise(E) = -\log_2(p(E)) \qquad (1) \qquad\qquad Surprise(E) = -\log_2(p(w1...wi|w1...wi-1)) \qquad (2)$$

Les modèles de coût cognitif syntaxique basés sur la surprise utilisent un parseur incrémental qui traite les phrases d'un texte mot par mot. À chaque nouveau mot, les règles du parseur peuvent être soit maintenues, soit écartées. Prenons comme exemple très simple la grammaire hors-contexte probabiliste présentée à la figure 1. Si la règle 1 est active, au sens où le parseur cherche à valider cette règle, alors les règles 2 et 3 sont toutes les deux disponibles. Si par la suite, le mot *le* (catégorie Det) est lu par le parseur, la règle 3 doit être écartée. La surprise du mot *le* est égale à la surprise de la masse de probabilité des règles maintenues. Plus formellement, la surprise d'un mot w_i est égale à la probabilité conditionnelle du préfixe de la phrase avec w_i inclus, sachant le préfixe de la phrase sans w_i (voir équation (2)). Alors, pour la grammaire de la figure 1, pour le mot *le* elle est égale à $-\log_2(0,7) = 0,51$. Si, dans un autre scénario, le parseur rencontrait le mot *John*, la situation demanderait le maintien de la règle 3 et l'écart de la règle 2. La surprise serait alors $-\log_2(0,3) = 1,74$. La surprise est donc plus haute quand on rencontre des structures moins fréquentes. La théorie de la surprise propose que le coût cognitif induit lors du traitement des phrases soit égal à la surprise (Hale, 2001).

L'adéquation des prédictions du modèle de surprise peut être illustrée par un exemple repris de Hale (2001) qui montre que le modèle simule correctement le fait que l'analyse par les humains est plus coûteuse pour une phrase relative objet que pour une phrase relative sujet. Dans la figure 2, reprise de Hale (2001), nous pouvons voir les prédictions du modèle pour ces deux types de phrase. À chaque mot, la surprise est calculée, en considérant les règles de la grammaire hors contexte représentée dans la figure 2. La différence entre la phrase relative sujet et la phrase relative objet se joue au niveau des règles respectives S[+R] → NP[+R] VP et S[+R] → NP[+R] S/NP [2]. Si le parseur rencontre une phrase relative, il ne sait pas s'il doit appliquer la première ou la deuxième règle jusqu'à un mot après *who*, qui est reconnu par la règle NP[+R] → *who*. Si l'on prend la phrase relative sujet de la figure 2, le moment où le parseur lit *saw*, la surprise est calculée par rapport à p(*the man who saw*|*the man who*). Or, cette probabilité est de $0,869 \cdot 1,0 \cdot 1,0$ et donne une surprise de $0,203$, car pour ajouter *saw* au préfixe de la phrase, on utilise les règles S[R+] → NP[+R] VP, VP → V NP et V → saw qui ont les probabilités respectives $0,869$, $1,0$ et $1,0$. Par contre, si après le préfixe *the man who*, le parseur

2. La phrase relative est indiquée par [+R] dans cette notation.

FIGURE 2 – Surprise pour les mots d'une phrase relative sujet et une phrase relative objet basée sur une grammaire hors contexte probabiliste. Figure reprise de Hale (2001).



Règle			probabilité
NP	$\rightarrow$	SPECNP NBAR	$0,33$
NP	$\rightarrow$	*you*	$0,33$
NP	$\rightarrow$	*me*	$0,33$
SPECNP	$\rightarrow$	DT	$1,0$
NBAR	$\rightarrow$	NBAR S[+R]	$0,5$
NBAR	$\rightarrow$	N	$0,5$
S	$\rightarrow$	NP VP	$1,0$
S[R+]	$\rightarrow$	NP[+R] VP	$0,869$
S[R+]	$\rightarrow$	NP[+R] S/NP	$0,131$
S/NP	$\rightarrow$	NP VP/NP	$1,0$
VP/NP	$\rightarrow$	V NP/NP	$1,0$
VP	$\rightarrow$	V NP	$1,0$
V	$\rightarrow$	*saw*	$1,0$
NP[+R]	$\rightarrow$	*who*	$1,0$
DT	$\rightarrow$	*the*	$1,0$
N	$\rightarrow$	*man*	$1,0$
NP/NP	$\rightarrow$	ϵ	$1,0$

lit le mot *you*, il doit calculer p(*the man who you*|*the man who*). Pour cela, il applique les règles S[R+] $\rightarrow$ NP[+R] S/NP, S/NP $\rightarrow$ NP VP/NP, et NP $\rightarrow$ *you* avec les probabilités respectives $0,131$, $1,0$ et $0,33$. La surprise pour la phrase relative objet est alors $-\log_2(0,131 \cdot 1,0 \cdot 0,33) = 4,527$.

La théorie de la surprise définit donc le coût cognitif en termes de probabilités des événements qui se manifestent lors de l'analyse syntaxique. Demberg & Keller (2008) ont trouvé qu'elle expliquait les temps de lectures sur le Dundee eye-tracking corpus (Kennedy *et al.*, 2003). En faisant un modèle linéaire mixte qui inclut aussi des facteurs de *bas niveau* comme, entre autres, la longueur de mots, la fréquence de mots et la position des mots sur la ligne, les auteurs ont trouvé que la surprise contribuait significativement à l'adéquation du modèle en s'appuyant critère d'information d'Akaike.

2.2 L'hypothèse de la réduction d'entropie

L'entropie est une propriété d'une distribution de probabilités. On pourrait dire que l'entropie est une mesure de l'incertitude d'une variable aléatoire. La notion d'entropie, telle que nous l'utilisons ici, vient de la théorie de l'information (Shannon & Weaver, 1949). L'entropie correspond au nombre de bits qui sont nécessaires en moyenne pour encoder le résultat de la variable aléatoire (Thomas &

Cover, 2006). L'entropie est maximale si tous les résultats de la variable aléatoire sont équiprobables. L'entropie $H(X)$ d'une variable aléatoire X a pour définition :

$$H(X) = -\sum_{j \in X} p(X = j) \cdot \log_2(p(X = j)) \tag{3}$$

La réduction d'entropie est calculée sur l'ensemble des parties droites des règles de réécriture des symboles non-terminaux dans une grammaire probabiliste hors contexte. Les grammaires récursives peuvent constituer un réel défi, car il y a une infinité de dérivations possibles. Mais, grâce au théorème de Grenander (1967), il est toujours possible de calculer l'entropie.

La réduction d'entropie est la différence entre l'entropie des règles de réécriture possibles avant de lire un mot de la phrase et les règles de réécriture possibles après avoir lu ce mot. L'écart des règles de réécriture change la distribution de probabilité et ceci résulte en une diminution d'entropie. Selon l'hypothèse de la réduction de l'entropie, l'entropie est une mesure de désambiguïsation (Hale, 2003, 2006). La désambiguïsation demande un coût cognitif, donc selon cette hypothèse, la réduction d'entropie est proportionnelle au coût cognitif. Cependant, il y a des théories qui ne prédisent pas le même rapport entre l'entropie et le coût cognitif. Linzen & Jaeger (2014) citent une théorie alternative qui prédit que la réduction d'entropie correspond non pas à une augmentation en coût cognitif, mais à une baisse de coût cognitif. Selon cette théorie, les situations où l'ambiguïté est maintenue demandent plus de coût cognitif que des situations où l'ambiguïté a été résolue. Malgré cette explication alternative, le modèle à effets mixtes linéaire de Linzen & Jaeger (2014) qui prédit le temps de lecture, mesuré par une tâche de lecture auto-segmentée (*self paced reading*), confirme l'hypothèse de la réduction d'entropie.

2.3 Les mesures intégrées

La théorie de la surprise et l'hypothèse de la réduction d'entropie ont formulé des mesures de coût cognitif liées à la structure syntaxique de la langue. Malgré le fait qu'elles capturent dans une certaine mesure des aspects lexicaux, si les symboles terminaux de la grammaire sont des mots formes, elles ne représentent pas de coût cognitif pour d'autres phénomènes linguistiques que la syntaxe. C'est pour cela que récemment, de nouvelles mesures ont été créées qui intègrent à la surprise ou à la réduction d'entropie d'autres facteurs. Ainsi Mitchell *et al.* (2010) proposent une mesure intégrée de la surprise venant de la syntaxe et de la sémantique compositionnelle des phrases. Pour un mot w_i, leur mesure est influencée par la surprise syntaxique pour w_i d'un côté, et de l'autre côté par la similarité sémantique pour deux vecteurs : un spécifique au mot w_i et l'autre spécifique au mot précédent w_{i-1}. Cette mesure intégrée leur permet d'obtenir un meilleur modèle pour les temps de lectures sur le Dundee Corpus (Kennedy *et al.*, 2003) que seule la surprise syntaxique ou la surprise sémantique.

Une autre mesure intégrée attire particulièrement notre attention, car il s'agit d'une mesure qui incorpore la tâche de la résolution de coréférence [3], qui est, en TAL en tout cas, liée à celle de la résolution d'anaphores. Dubey *et al.* (2013) ont inventé une mesure qui, selon eux, peut capturer le coût cognitif des phénomènes de discours en plus de la structure syntaxique. Avec un outil de

3. La résolution de la coréférence est une tâche consistant à repérer dans le texte toutes les références se rapportant à la même entité du discours. Ainsi dans la phrase *Hier, **[Alex]**, **[le roi des Pays-Bas]**, **[s']**est cogné le pied.* tous les éléments en gras entre crochets appartiennent à une chaîne de coréférence.

coréférence très basique, ils décident pour chaque syntagme nominal dans un texte si ce syntagme représente une nouvelle entité de discours, ou une entité déjà connue. Si l'outil prédit qu'il s'agit d'une nouvelle entité, le coût prédit sera plus haut. Cela veut dire que ce modèle prédit que les nouvelles entités de discours demandent plus de coût cognitif que les entités de discours connues. En comparant la variance dans les temps de lecture dans le Dundee eye-tracking corpus (Kennedy *et al.*, 2003) expliquée par un modèle qui utilise seulement la surprise venant de la syntaxe et celle expliquée par leur modèle de coût intégré, ils ont trouvé que la mesure intégrée expliquait mieux la variance : en utilisant les modèles mixtes, le modèle qui incluait la mesure intégrée obtenait un score de vraisemblance significativement plus élevée.

À notre connaissance, Dubey *et al.* (2013) représente la seule étude où la coréférence est utilisée pour prédire le coût cognitif sur corpus. Bien qu'il soit encourageant de voir qu'il y a un travail qui tente de traiter le problème du coût cognitif venu des facteurs de discours, nous pensons que le modèle proposé par Dubey *et al.* (2013) est loin d'être complet. Le modèle ne prend rien d'autre en compte que le résultat de l'outil de la coréférence qui prédit si un syntagme nominal représente une nouvelle entité de discours, ou une ancienne. L'objectif de notre modèle du coût cognitif des anaphores est de prédire le coût cognitif à partir de multiples facteurs. Notre ambition est de développer un modèle qui va plus loin que la simple heuristique d'augmenter le coût si un syntagme nominal est nouveau dans le discours. Dans le prochain paragraphe, nous décrivons notre première tentative pour construire un tel modèle.

3 Un modèle probabiliste de coût cognitif basé sur l'entropie et un algorithme par couples de mentions

Dans cette section, nous présentons notre mesure de coût cognitif pour les anaphores. Pour cette première tentative, nous avons décidé de nous restreindre aux pronoms personnels de la troisième personne [4]. Nous avons fait ce choix pour deux raisons. La première raison est que la littérature psycholinguistique traitant des anaphores se concentre surtout les pronoms personnels de troisième personne. La deuxième raison est qu'en faisant des expériences préliminaires, nous avons trouvé que le système de résolution de coréférence était moins performant sur la résolution des pronoms que sur les noms communs et les noms propres. En éliminant ces deux dernières catégories, le système classifie beaucoup mieux les pronoms, car la tâche de classification est plus homogène et cela permet au système d'apprentissage automatique de mieux apprendre.

Pour modéliser le coût, notre modèle utilise des probabilités venues d'un algorithme par couples de mentions (Soon *et al.*, 2001). Ce premier système de résolution d'anaphores est entraîné sur le corpus ANCOR Centre (Muzerelle *et al.*, 2014), un corpus oral de la langue française qui compte 488 000 mots et qui est en l'occurrence le seul grand corpus annoté en coréférence disponible pour le français. Désoyer *et al.* (2015) ont réussi à entraîner un système par couples de mentions fonctionnel de la résolution de coréférence sur ce corpus, ce qui nous a incité à utiliser ce corpus pour notre modèle.

Dans cette section nous expliquerons d'abord le fonctionnement de l'algorithme par couples de mentions, pour expliquer ensuite comment on l'utilise pour créer une mesure de coût cognitif. Puis nous montrerons comment nous avons implémenté notre mesure pour prédire le cout cognitif sur les pronoms dans le corpus ANCOR Centre. Cette section se termine par une visualisation de la

4. 'il', 'ils', 'elle', 'elles', 'le', 'la', 'l", 'lui', 'les', 'eux'

prédiction de coût cognitif et une discussion des résultats.

3.1 L'algorithme par couples de mentions

Un des algorithmes les plus connus de résolution de coréférence par une approche statistique est l'algorithme par couples de mentions de Soon *et al.* (2001). Nous avons choisi d'utiliser cet algorithme, parce qu'il peut s'adapter facilement à notre problématique de résolution de pronoms de troisième personne. De plus, cet algorithme est toujours au cœur de des systèmes du niveau de l'état de l'art comme BART (Broscheit *et al.*, 2010) et a le mérite d'être simple à implémenter. L'algorithme utilise la classification supervisée pour décider si deux descriptions référentes, désormais appelées *mentions*, sont coréférentes ou pas. Le classifieur est entraîné avec des couples de mentions coréférentes et non coréférentes, pour apprendre une représentation de ces deux types de couples. Le but de l'algorithme est de former des chaînes de coréférence (voir la note 3). Pour ajouter une mention m_j du document à une chaîne de coréférence, des couples avec toutes les mentions précédentes m_i avec $i \in [0, 1, ..., j-1]$ dans le document sont formés : $\langle (m_0, m_j), (m_1, m_j), ..., (m_{j-1}, m_j) \rangle$ [5]. Pour chaque couple (m_i, m_j) on applique le classifieur. Si le classifieur est probabiliste, on obtient aussi $p((m_i, m_j) = \mathit{coref})$. Ensuite, parmi les mentions qui ont une probabilité d'être coréférente à m_j supérieure à 50 %, il faut choisir l'antécédent. Soon *et al.* (2001) utilisent une heuristique qui dit que l'antécédent est la mention la plus proche pour laquelle le classifieur a attribué une probabilité supérieure à $0,5$. Dans notre travail, nous utiliserons l'algorithme par couples de mentions probabiliste, sauf que pour prédire le coût cognitif nous n'avons pas besoin de l'heuristique de décision et nous faisons seulement la résolution des pronoms. C'est-à-dire que les mentions m_j à résoudre ne sont que des pronoms, alors que les mentions m_i — parmi lesquelles on recherche l'antécédent — contiennent tous les types de mentions (syntagmes nominaux, pronoms, etc).

3.2 L'entropie en tant que mesure de difficulté des pronoms

Nous avons basé notre mesure de coût cognitif sur l'entropie. Comme décrit ci-dessus dans le paragraphe 2.2, l'entropie peut être considérée comme une mesure d'ambiguïté. Dans notre modèle, nous établissons la difficulté des anaphores en comparant une situation où l'ambiguïté et l'entropie sont maximales à une situation où une tentative de désambiguïser a été faite. La première situation correspond au cas où le système (le solveur par couples de mentions) attribue une probabilité de 0,5 à chaque antécédent candidat d'un pronom. La deuxième situation correspond au cas où le système calcule proprement toutes les probabilités pour les couples de mentions qui incluent le pronom à résoudre m_j et les antécédents candidats m_i avec $i \in [0, 1, ..., j-1]$.

Si le système n'a pas de difficulté à trouver pour chaque candidat antécédent s'il est le antécédent du pronom ou non, le système ne donnera que des probabilités proches de 1 et de 0. Ces probabilités fortes et faibles indiquent que le système est sûr de ses décisions. Nous pouvons par la suite étudier la distribution jointe sur les probabilités que le système a donné pour les couples de mentions (m_i, m_j). Nous faisons cela en considérant chaque couple de mentions comme une variable aléatoire qui peut prendre les valeurs 1 et 0, 1 dans le cas où $(m_i, m_j) = \mathit{coref}$ et 0 dans le cas où $(m_i, m_j) = \neg \mathit{coref}$. Les sorties 1 et 0 ont chacune une probabilité associée. C'est sur l'ensemble des sorties possibles

5. Bien qu'il soit possible de chercher l'antécédent depuis le début du texte, souvent la recherche est restreinte à l'historique récente, par exemple aux dix dernières phrases qui précèdent la mention à résoudre.

FIGURE 3 – Probabilités pour trois couples d'un pronom m_j avec trois candidats antécédents.

m_i	m_j	$p(m_i, m_j = coref)$
m_1	m_4	0,7
m_2	m_4	0,2
m_3	m_4	0,1

FIGURE 4 – La probabilité jointe sur les probabilités de la figure 3.

Couple (m_i, m_j)	1	2	3	4	5	6	7	8
(1,4)	1	1	1	0	0	0	0	1
(2,4)	1	1	0	0	0	1	1	0
(3,4)	1	0	0	0	1	1	0	1
probabilité	0,014	0,126	0,504	0,216	0,024	0,006	0,054	0,056

du système de ces variables aléatoires, la distribution jointe, que nous pouvons calculer l'entropie [6]. Cette entropie peut par la suite être comparée à l'entropie maximale pour cette distribution. Si une mention m_j est facile à résoudre pour le système, l'entropie doit être basse quand elle est comparée à l'entropie maximale. Nous pouvons donc formuler une mesure de difficulté de résolution de pronom de la façon suivante : la difficulté est le ratio entre l'entropie de la probabilité jointe de tous les couples (m_i, m_j) avec $i \in [0, 1, ..., j-1]$ et son entropie maximale.

Regardons de plus près un exemple concret pour calculer la difficulté d'un pronom. Admettons qu'il y ait un pronom m_4 avec trois mentions précédentes et les probabilités de la figure 3. L'ensemble des sorties possibles du système est représenté à la figure 4. Il y a huit sorties de système possibles pour les différents couples de mentions. La distribution probabiliste sur ces huit sorties possibles est la distribution jointe sur les probabilités de la figure 3. La difficulté du pronom m_4 serait :

$$
\begin{aligned}
H_{distribution\ jointe\ de\ tous\ les\ couples\ m_4}/H_{max} = & \\
& (0,014 \cdot \log_2(0,014)+ \\
& 0,126 \cdot \log_2(0,126)+ \\
& 0,504 \cdot \log_2(0,504)+ \\
& 0,216 \cdot \log_2(0,216)+ \\
& 0,024 \cdot \log_2(0,024)+ \\
& 0,006 \cdot \log_2(0,006)+ \\
& 0,054 \cdot \log_2(0,054)+ \\
& 0,056 \cdot \log_2(0,056)) \\
& / (8 \cdot 0,125 \cdot \log_2(0,125)) \\
= & 2,072/3 \\
= & 0,69
\end{aligned}
$$

3.3 Mise en œuvre

Dans ce paragraphe nous détaillerons comment nous avons implémenté le modèle. Le solveur par couples de mentions est un classifieur qui utilise la régression logistique. Pour son implémentation, la bibliothèque d'apprentissage statistique *scikit-learn* en langage Python (Pedregosa *et al.*, 2011) a été utilisée. La régression logistique a été choisie pour deux raisons. D'abord elle est connue pour retourner des probabilités calibrées. Cela veut dire qu'à chaque fois que le classifieur attribue une

6. Notre système fait la simplification que toutes les probabilités $p((m_i, m_j)) = coref$ sont indépendantes, bien que cela soit probablement faux. Mais, l'architecture du système couples de mentions permet de voir chaque couple (m_i, m_j) comme indépendant, car pendant le calcul de $p((m_i, m_j) = coref)$ les autres couples ne sont pas pris en considération.

probabilité de 0,60, le classifieur donne dans 60% des cas la bonne réponse. D'autres algorithmes d'apprentissage machine, par exemple *les séparateurs à vaste marge* (SVM), ont une distribution plus dissymétrique [7]. Deuxièmement, la régression logistique retourne des coefficients qui représentent l'importance de chaque facteur que l'algorithme prend en compte pour faire ces classifications et nous fournit une importante indication sur la façon dont différents facteurs jouent un rôle plus proéminent dans la résolution des anaphores que d'autres. Pour nos expériences, nous avons divisé le corpus ANCOR Centre en un corpus d'entraînement, de développement et de test. Nous avons suivi le découpage de Désoyer *et al.* (2015), qui proposait des tailles respectives de 60%, 20% et 20%. Les tailles de ces sous-corpus permettaient d'entraîner, développer et tester convenablement le modèle.

3.3.1 Entraînement

Notre classifieur a été entraîné avec des vecteurs de traits qui chacun représentent un couple de deux mentions. Ces mentions sont soit coréférentes, soit non-coréférentes. Les exemples d'entraînement ont été sélectionnés selon la méthode présentée dans Soon *et al.* (2001), c'est-à-dire : les exemples positifs sont les couples des pronoms anaphoriques et de leur antécédent les plus proche, les exemples négatifs sont les couples formés de ces mêmes pronoms et de toutes les mentions placées entre ces pronoms et leur antécédent le plus proche [8]. De notre corpus d'entraînement, nous récupérons ainsi 4322 couples positifs (coréférents) et 10015 couples négatifs (non-coréférents). Le classifieur apprend une représentation de ces deux types de couples en forme de vecteurs de onze traits. Dans la figure 5, la nature de ces traits est spécifiée. Nous avons trouvé les valeurs de ces traits dans l'annotation du corpus ANCOR Centre.

FIGURE 5 – Les traits de classification sont présentés avec leurs coefficients. La grandeur des coefficients des traits indique leur importance. Pour les sept premiers traits qui sont booléens, un coefficient négatif indique que si ce trait est vrai, la probabilité que le couple est coréférent est plus basse et un coefficient positif indique l'inverse. Les quatre derniers traits de distance sont continus. Pour ceux qui ont des coefficients négatifs, plus leur valeur est élevée, moins il est probable que le couple soit coréférent. L'inverse est vrai pour les coefficients positifs.

Coefficient	Trait de classification
$-6,934$	ordonnée à l'origine
$-1,158$	la première mention se trouve à l'intérieur d'un syntagme prépositionnel
$-0,203$	la deuxième mention se trouve à l'intérieur d'un syntagme prépositionnel
$-1,448$	la première mention commence par un déterminant démonstratif
$-0,737$	la première mention est un nom indéfini
$2,235$	les deux mentions partagent le même nombre
$2,648$	les deux mentions partagent le même genre
$1,366$	les deux mentions partagent la même partie de discours
$-0,031$	nombre de mentions entre les deux mentions du couple
$-0,003$	distance d'édition (Lehvenstein) entre les chaînes de caractères des deux mentions
$0,002$	distance en nombre de mots entre la première mention et le début du tour de parole
$0,001$	distance en nombre de mots entre la deuxième mention et le début du tour de parole

FIGURE 6 – Les résultats du classifieur sur les sous-parties du corpus ANCOR Centre.

Sous-corpus	Taille	Pronoms à Résoudre	Précision	Rappel	F-mesure
Entraînement	60%	4322	0,76	0,84	0,80
Développement	20%	1389	0,78	0,82	0,80
Test	20%	1245	0,76	0,81	0,78

7. http://scikit-learn.org/stable/modules/calibration.html

8. Bien que cette méthode puisse induire un biais par rapport aux traits de classification de distance, elle s'est avéré le plus efficace lors de nos premières expérimentations.

3.3.2 Évaluation du solveur

Nous avons évalué le classifieur sur les corpus d'entraînement développement et de test. Pour chaque pronom anaphorique dans les corpus, le classifieur devait trouver le référent de l'anaphore parmi les dix mentions précédentes [9]. Dans la figure 6, on peut voir la performance du classifieur en termes de précision, rappel et F-mesure par sous-corpus [10].

3.3.3 Visualisation des résultats

Dans la figure 7 nous pouvons voir les différents scores de difficulté d'une grande partie des pronoms d'un texte dans le corpus. Nous pouvons voir qu'il y a beaucoup de variation entre les scores de difficulté de différents pronoms. Une deuxième visualisation nous permet ensuite de voir pourquoi le score d'une anaphore m_j est haut ou bas. Pour chaque candidat antécédent m_i, nous avons pris le texte qui précédait dans le corpus et pour les candidats antécédents nous avons indiqué les scores que le solveur d'anaphores attribuait aux couples (m_i, m_j). Dans la figure 8, nous pouvons voir respectivement une anaphore qui a été jugée difficile par notre système et une qui a été jugée beaucoup plus facile avec sept antécédents candidats.

3.4 Discussion

Nous pouvons voir que les scores pour les antécédents candidats se rapprochent beaucoup plus de 0 et de 1 pour l'anaphore facile. Pour l'anaphore difficile nous pouvons voir qu'il y a beaucoup de scores intermédiaires. Pour l'anaphore facile, il y a beaucoup d'antécédents candidats qui ne correspondent pas en genre en en nombre, en plus, beaucoup d'entre eux se trouvent dans une phrase prépositionnelle comme *de l'entreprise, sur ce bâtiment, de toute la faculté, des sciences*. Comme notre système pénalise cela, ce que nous pouvons voir dans la figure 5 par les coefficients négatifs, ces éléments ont un score très bas. Pour l'anaphore difficile par contre, il y a tout d'abord une ambiguïté pour les candidats référents : *le latin* et *l'enfant*. Puis, excepté dans *du latin*, le trait qui pénalise la position à l'intérieur d'un syntagme prépositionnel ne joue pas de rôle pour les antécédents candidats de cette anaphore.

Mais, autant de difficulté pour l'anaphore difficile ne semble tout de même pas justifiée : une partie de la difficulté est ajoutée par l'architecture de notre système. Il faut d'abord remarquer que beaucoup de candidats référents font partie d'une même chaîne de coréférence, en l'occurrence *l'enfant, lui-même, l'enfant, lui, lui*. La probabilité d'attacher le dernier *lui* aux éléments de cette chaîne n'est pas extrêmement élevée, elle est à environ 0,6. Le problème est que cette probabilité est comptée cinq fois, car notre système n'a pas la notion que ces cinq éléments font partie d'une même chaîne de coréférence. L'impact de la probabilité intermédiaire de 0,6 est donc amplifié par une longue chaîne de coréférence, ce qui fait croître l'entropie d'une façon indésirable. Il faut aussi remarquer que le système de résolution d'anaphores donne rarement un score supérieur à 0,8 à un couple (m_i, m_j). Cela est en lien avec la calibration de notre solveur : comme la précision du système est d'environ 0,7 à 0,8 (voyez la figure 6), le système donne des probabilités autour de cette valeur aux couples qu'il

9. Cette restriction correspond à la limitation de l'historique de recherche souvent appliquée pour l'algorithme de couples de mentions. Comme nous n'avons pas de phrases dans le corpus, nous avons exprimée l'historique en nombre de mentions.

10. La taille des sous-corpus correspond au pourcentage de nombre de textes contenus dedans, c'est pour cela que dans le corpus d'entraînement il n'y a pas exactement 60% des pronoms.

FIGURE 7 – Scores de difficulté de différentes anaphores au sein d'un même document. Les deux anaphores entourées sont celles de la figure 8. On a identifié les anaphores par leur mot forme et leur identifiant dans le corpus ANCOR Centre

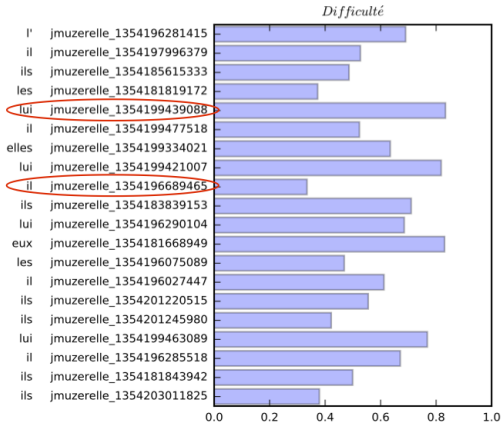
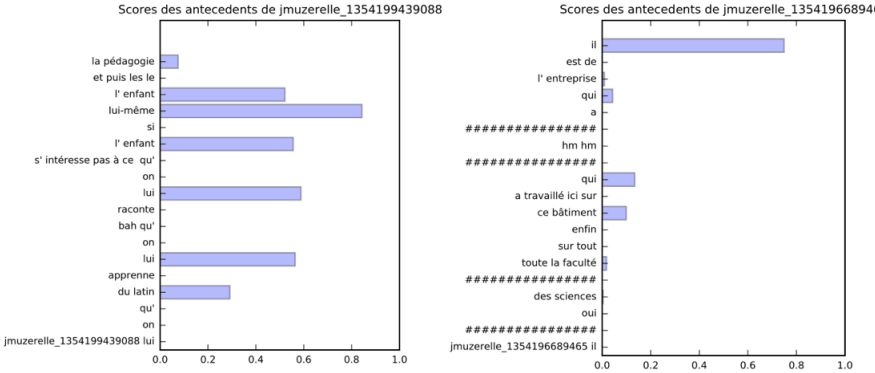


FIGURE 8 – Une anaphore difficile (à gauche) et une anaphore facile (à droite). Les tours de paroles sont indiquées par des croisillons.



classe coréférents. Ainsi, quand il y a plusieurs réponses positives du classifieur pour les couples (m_i, m_j), l'entropie augmente. Cela n'est pas problématique si les différentes m_i pour lesquelles le classifieur retourne positif sont vraiment des candidats antécédents différents, mais quand elles appartiennent en fait à une même chaîne de coréférence, cela ne devrait pas être le cas. Pour la suite de notre travail, il est donc important que nous intégrions une notion d'appartenance à une chaîne de coréférence dans le modèle.

4 Conclusion

Inspirés par les modèles syntaxiques de coût cognitif, nous avons fait un premier pas vers un modèle de coût cognitif pour la résolution d'anaphores. Notre système probabiliste par couples de mentions a

le mérite de nous donner les probabilités de coréférence pour les couples composés d'un pronom et ses candidats antécédents, que nous pouvons utiliser pour mesurer l'entropie. Le ratio de cette entropie et de l'entropie maximale nous donne une mesure de coût cognitif de la résolution d'anaphores. Les premières visualisations nous ont permis de constater que notre modèle prédit un coût cognitif qui varie selon le contexte de l'anaphore. Néanmoins, nous avons aussi constaté que dans des cas où l'anaphore appartient à une longue chaîne de coréférence, le modèle prédit trop de coût. Nos premiers résultats nous donnent des indications sur la suite de notre recherche que nous préciserons ci-dessous.

Dans le paragraphe 3.4, nous avons vu que le système ne disposait pas de notion d'appartenance à une chaîne et que cela ajoutait une difficulté pour les chaînes de coréférence longues. Il serait donc nécessaire de modéliser la résolution d'anaphores comme un rattachement à une entité de discours, plutôt qu'à un seul candidat référent. Nous réfléchissons en ce moment à un modèle qui ne se baserait pas sur une probabilité jointe des scores de couples de mentions, mais sur la distribution de probabilité sur les chaînes de coréférence possibles au sein d'un document, comme dans les travaux de Luo *et al.* (2004).

Un deuxième changement à faire est d'ajouter plus de contraintes syntaxiques au modèle, car elles jouent un rôle important dans la résolution d'anaphores aussi bien en TAL qu'en psycholinguistique (Frazier, 2015). La raison pour laquelle il n'y a pas de traits syntaxiques dans le modèle actuel est qu'il est développé sur un corpus oral, faute de corpus écrit et annoté en coréférence pour le français, et qui ne peut pas être analysé automatiquement pour la syntaxe. Le prochain étape de notre projet sera de continuer pour la langue anglaise, pour laquelle des corpus annotés en coréférence du langage écrit sont disponibles.

Un troisième point est qu'il faudrait réfléchir à la façon d'intégrer certains facteurs connus en psycholinguistique, mais pas (souvent) utilisés en TAL, comme par exemple la structure de l'information, la causalité implicite, les connecteurs de discours ou les connaissances du monde. Il faudrait trouver un moyen de représenter ces aspects sous forme de traits de classification. L'utilisation de ce type de traits est très important pour notre projet de recherche, car ce ne sera qu'en implémentant ce genre de traits que nous pouvons évaluer les différentes théories psycholinguistiques, comme par exemple la théorie de l'accessibilité (Ariel, 1990) ou la théorie du centrage (Grosz *et al.*, 1995).

Un dernier point, et pas le moindre, est d'établir une méthode pour confronter les prédictions de notre modèle à des mesures de coût cognitif des humains. Nous avons l'intention d'utiliser des temps de lecture sur corpus. Pour le français nous comptons utiliser le *French Treebank* (Abeillé *et al.*, 2003), qui a une annotation en temps de lecture par occulométrie. Pour nos futures expériences en anglais, nous comptons utiliser le *Dundee Corpus* (Kennedy *et al.*, 2003), qui est annoté de la même manière. Nous devons trouver une façon d'associer un temps de lecture aux anaphores. Ceci peut constituer une tâche difficile, car les mots courts ne sont pas souvent fixés par des humains quand ils lisent. Il faudra donc élaborer une façon de définir les régions de lecture pour les anaphores.

Remerciements

Nous remercions les deux relecteurs Laurence Longo et Thomas François pour leurs suggestions et commentaires, ainsi que nos collègues Maximin Coavoux et Sarah Beniamine. Ce travail de thèse a été encadré par Pascal Amsili et Adeline Nazarenko et est soutenu par l'école doctorale Frontières du Vivant de l'USPC, ainsi que par le Labex EFL (ANR-10-LABX-0083).

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for french. In *Treebanks*, p. 165–187. Springer.

ARIEL M. (1990). *Accessing noun-phrase antecedents (rle linguistics b : Grammar)*. Routledge.

BROSCHEIT S., POESIO M., PONZETTO S. P., RODRIGUEZ K. J., ROMANO L., URYUPINA O., VERSLEY Y. & ZANOLI R. (2010). Bart : A multilingual anaphora resolution system. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, p. 104–107 : Association for Computational Linguistics.

CARAMAZZA A., GROBER E., GARVEY C. & YATES J. (1977). Comprehension of anaphoric pronouns. *Journal of verbal learning and verbal behavior*, **16**(5), 601–609.

COLONNA S., SCHIMKE S. & HEMFORTH B. (2015). Different effects of focus in intra-and inter-sentential pronoun resolution in german. *Language, Cognition and Neuroscience*, **30**(10), 1306–1325.

DEMBERG V. & KELLER F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, **109**(2), 193–210.

DÉSOYER A., LANDRAGIN F., TELLIER I., LEFEUVRE A. & ANTOINE J.-Y. (2015). Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ancor. *Traitement Automatique des Langues*, **55**(2), 97–121.

DUBEY A., KELLER F. & STURT P. (2013). Probabilistic modeling of discourse-aware sentence processing. *Topics in cognitive science*, **5**(3), 425–451.

FRANK S. L. (2013). Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in cognitive science*, **5**(3), 475–494.

FRAZIER L. (2015). Squib : co-reference and adult language comprehension. *Revista LinguiStica*, **8**(2).

GRENANDER U. (1967). *Syntax-controlled probabilities*. Division of Applied Mathematics, Brown University.

GROSZ B. J., WEINSTEIN S. & JOSHI A. K. (1995). Centering : A framework for modeling the local coherence of discourse. *Computational linguistics*, **21**(2), 203–225.

HALE J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, p. 1–8 : Association for Computational Linguistics.

HALE J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, **32**(2), 101–123.

HALE J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, **30**(4), 643–672.

KEHLER A. & ROHDE H. (2013). A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, **39**(1-2), 1–37.

KENNEDY A., HILL R. & PYNTE J. (2003). The dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

KOORNNEEF W. A. (2008). *Eye-catching Anaphora*. PhD thesis, Universiteit Utrecht.

LEVY R. (2008). Expectation-based syntactic comprehension. *Cognition*, **106**(3), 1126–1177.

LINZEN T. & JAEGER T. F. (2014). Investigating the role of entropy in sentence processing. In *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics*, p. 10–18.

LUO X., ITTYCHERIAH A., JING H., KAMBHATLA N. & ROUKOS S. (2004). A mention-synchronous coreference resolution algorithm based on the bell tree. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, p. 135–143 : Association for Computational Linguistics.

MITCHELL J., LAPATA M., DEMBERG V. & KELLER F. (2010). Syntactic and semantic factors in processing difficulty : An integrated measure. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 196–206 : Association for Computational Linguistics.

MUZERELLE J., LEFEUVRE A., SCHANG E., ANTOINE J.-Y., PELLETIER A., MAUREL D., ESHKOL I. & VILLANEAU J. (2014). Ancor centre, a large free spoken french coreference corpus : description of the resource and reliability measures. In *LREC'2014, 9th Language Resources and Evaluation Conference.*, p. 843–847.

NIEUWLAND M. S. & VAN BERKUM J. J. (2008). The interplay between semantic and referential aspects of anaphoric noun phrase resolution : Evidence from erps. *Brain and Language*, **106**(2), 119–131.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

SHANNON C. E. & WEAVER W. (1949). The mathematical theory of communication (urbana, il.

SHELDON R. *et al.* (1998). *A first course in probability*. Prentice-Hall.

SOON W. M., NG H. T. & LIM D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, **27**(4), 521–544.

THOMAS J. A. & COVER T. (2006). *Elements of information theory*. Wiley New York, 2 edition.