Cognitive Computational Models of Pronoun Resolution

Olga Seminck

université PARIS DIDEROT PARIS 7

USPC Université Sorbonne Paris Cité

Thèse de doctorat

de l'Université Sorbonne Paris Cité

Préparée à l'Université Paris Diderot

Ecole doctorale Frontières du Vivant (ED 474)

Laboratoire de Linguistique Formelle & Laboratoire d'Informatique de Paris Nord

Cognitive Computational Models of Pronoun Resolution

Olga Seminck

Thèse de doctorat de Linguistique

Dirigée par Pascal Amsili et Adeline Nazarenko

Présentée et soutenue publiquement à Paris le 23 novembre 2018

Rapporteur & Président : Prévot, Laurent, Professeur, Université d'Aix-Marseille

Rapporteur : Villavicencio, Aline, Lecturer, University of Essex

Examinateur : Colonna, Saveria, Maître de Conférences HDR, Université Paris 8 Vincennes – Saint-Denis

Examinateur : Demberg, Vera, Professeur, Universität des Saarlandes

Directeur : Amsili, Pascal, Maître de Conférences HDR, Université Paris Diderot

Co-directeur : Nazarenko, Adeline, Professeur, Université Paris 13

Résumé court

Modèles cognitifs et computationnels de la résolution des pronoms

La résolution des pronoms est le processus par lequel un pronom anaphorique est mis en relation avec son antécédent. Les humains en sont capables sans efforts notables en situation normale. En revanche, les systèmes automatiques ont une performance qui reste loin derrière, malgré des algorithmes de plus en plus sophistiqués, développés par la communauté du Traitement Automatique des Langues.

La recherche en psycholinguistique a montré à travers des expériences qu'au cours de la résolution de nombreux facteurs sont pris en compte par les locuteurs. Une question importante se pose : comment les facteurs interagissent et quel poids faut-il attribuer à chacun d'entre eux ? Une deuxième question qui se pose alors est comment les théories linguistiques de la résolution des pronoms incorporent tous les facteurs.

Nous proposons une nouvelle approche à ces problématiques : la simulation computationnelle de la charge cognitive de la résolution des pronoms. La motivation pour notre approche est double : d'une part, l'implémentation d'hypothèses par un système computationnel permet de mieux spécifier les théories, d'autre part, les systèmes automatiques peuvent faire des prédictions sur des données naturelles comme les corpus de mouvement oculaires. De cette façon, les modèles computationnels représentent une alternative aux expériences classiques avec des items expérimentaux construits manuellement.

Nous avons fait plusieurs expériences afin d'explorer les modèles cognitifs computationnels de la résolution des pronoms. D'abord, nous avons simulé la charge cognitive des pronoms en utilisant des poids de facteurs de résolution appris sur corpus. Ensuite, nous avons testé si les concepts de la Théorie de l'Information sont pertinents pour prédire la charge cognitive des pronoms. Finalement, nous avons procédé à l'évaluation d'un modèle psycholinguistique sur des données issues d'un corpus enrichi de mouvements oculaires.

Les résultats de nos expériences montrent que la résolution des pronoms est en effet multi-factorielle et que l'influence des facteurs peut être estimée sur corpus. Nos résultats montrent aussi que des concepts de la Théorie de l'Information sont pertinents pour la modélisation des pronoms.

Nous concluons que l'évaluation des théories sur des données de corpus peut jouer un rôle important dans le développement de ces théories et ainsi amener dans le futur à une meilleure prise en compte du contexte discursif.

Mots clés : pronom, anaphore, coréférence, résolution, modèle mixte, occulométrie, temps de lecture, corpus, métrique de coût cognitif, Théorie de l'Information

Abstract

Cognitive Computational Models of Pronoun Resolution

Pronoun resolution is the process in which an anaphoric pronoun is linked to its antecedent. In a normal situation, humans do not experience much cognitive effort due to this process. However, automatic systems perform far from human accuracy, despite the efforts made by the Natural Language Processing community.

Experimental research in the field of psycholinguistics has shown that during pronoun resolution many linguistic factors are taken into account by speakers. An important question is thus how much influence each of these factors has and how the factors interact with each-other. A second question is how linguistic theories about pronoun resolution can incorporate all relevant factors.

In this thesis, we propose a new approach to answer these questions: computational simulation of the cognitive load of pronoun resolution. The motivation for this approach is two-fold. On the one hand, implementing hypotheses about pronoun resolution in a computational system leads to a more precise formulation of theories. On the other hand, robust computational systems can be run on uncontrolled data such as eye movement corpora and thus provide an alternative to hand-constructed experimental material.

In this thesis, we conducted various experiments. First, we simulated the cognitive load of pronouns by learning the magnitude of impact of various factors on corpus data. Second, we tested whether concepts from Information Theory were relevant to predict the cognitive load of pronoun resolution. Finally, we evaluated a theoretical model of pronoun resolution on a corpus enriched with eye movement data.

Our research shows that multiple factors play a role in pronoun resolution and that their influence can be estimated on corpus data. We also demonstrate that the concepts of Information Theory play a role in pronoun resolution. We conclude that the evaluation of hypotheses on corpus data enriched with cognitive data —- such as eye movement data — play an important role in the development and evaluation of theories. We expect that corpus based methods will lead to a better modelling of the influence of discourse structure on pronoun resolution in future work.

Key words: pronoun, anaphora, coreference, resolution, mixed effects model, eye-tracking, reading times, corpus, cognitive cost metric, Information Theory

Résumé long

Modèles cognitifs et computationnels de la résolution des pronoms

Cette thèse porte sur les modèles cognitifs et computationnels : des modèles qui ont pour but de simuler le traitement cognitif du langage par les humains. Nous cherchons à développer ces modèles pour la résolution des pronoms. Bien que la résolution des pronoms ne soit pas un nouveau sujet de recherche — d'innombrables études ont été effectuées — le sujet n'a presque jamais été étudié sous l'angle des modèles computationnels et cognitifs. Nous sommes d'avis que ces modèles ont un grand potentiel de renforcer la recherche sur la résolution des pronoms. Le but de cette thèse est donc de développer de tels modèles.

L'objectif des modèles cognitifs et computationnels est de simuler le comportement des humains au moyen de programmes informatiques. Un tel programme, confronté à des stimuli linguistiques, doit fournir une sortie qui soit aussi proche que possible de celle que les humains produisent. La sortie du programme informatique est calculée en fonction d'une modélisation d'un phénomène linguistique présent dans ces stimuli. En comparant cette sortie aux réponses humaines face aux mêmes stimuli, la plausibilité des hypothèses sous-jacentes à la modélisation peut être évaluée.

Il est important de noter que seule la plausibilité de la théorie à la base de la modélisation peut être évaluée de cette façon. En effet, le fait qu'un programme informatique produise des réponses similaires à celles des humains n'est pas une preuve que le programme ait modélisé le traitement cognitif d'une manière adéquate. Inversement, un échec de simulation ne signifie pas non plus que les hypothèses sous-jacentes soient fausses. Il se pourrait simplement que le programme manque de précision et échoue malgré le fait d'être basé sur une théorie correcte.

Malgré le fait que la modélisation cognitive computationnelle ne constitue pas à elle seule un moyen d'obtenir des preuves solides pour une théorie, elle est tout de même utile. Tout d'abord, ces modèles donnent une implémentation à des théories. Ceci semble trivial, mais il n'en est rien. Quand on implémente une théorie, toute prédiction faite par celle-ci doit être traduite en actions par le programme informatique. Cela implique que l'on doit réfléchir sur tous les détails de la théorie parce que le programme doit savoir à tout moment quelle action il faut entreprendre.

Une deuxième application utile des modèles cognitifs et computationnels est la comparaison des théories. Quand deux théories sont implémentées, on peut mesurer laquelle simule le mieux le comportement humain. Quand l'implémentation de la théorie A produit des résultats plus proches aux réponses humaines que la théorie B, on peut conclure que la théorie A est plus plausible que la théorie B. Cette façon de comparer les théories est d'usage général dans le domaine de la modélisation cognitive et computationnelle.

L'évaluation des modèles cognitifs et computationnels peut aller de pair avec une évaluation sur corpus. La plupart des théories sur le traitement cognitif de la langue s'appuient sur des résultats d'études psycholinguistiques. Les stimuli utilisés dans ces études ont été conçus pour que des phénomènes linguistiques autres que ceux à l'étude soient exclus. Par contre, dans un environnement naturel, de nombreux phénomènes jouent un rôle en même temps et interagissent. Utiliser un corpus — à condition que celui-ci n'ait pas été construit pour contenir ou exclure spécifiquement

certains phénomènes linguistiques — est un moyen pour étudier à quel point une théorie fait des prédictions correctes sur des données naturelles. Les études sur des corpus offrent donc un moyen d'évaluer la robustesse des théories. Si une théorie peut toujours faire des prédictions adéquates malgré l'influence des phénomènes linguistiques qui ne sont pas l'objet d'étude, on peut conclure qu'elle est robuste.

Une dernière raison pour laquelle les modèles cognitifs et computationnels sont utiles est qu'ils peuvent être une inspiration pour l'intelligence artificielle (IA), dont le but en général est de développer des programmes capables d'effectuer des tâches aussi bien, voir mieux, que les humains. On s'intéresse ici à la branche de l'IA qui porte sur la linguistique, donc au domaine du Traitement Automatique des Langues (TAL). En TAL, des programmes sont développés pour des tâches telles que la traduction automatique, l'analyse grammaticale des phrases et le raisonnement automatique. Si une meilleure modélisation du traitement cognitif peut être atteinte grâce aux modèles cognitifs et computationnels, elle devrait donc être une source d'inspiration pour le domaine du TAL.

Les modèles cognitifs et computationnels de la résolution des pronoms sont un sujet d'étude récent. La littérature sur les modèles cognitifs et computationnels sur le traitement syntaxique et lexical se développe, mais il n'y que très peu de modèles traitant d'autres phénomènes linguistiques. Quant à la résolution des pronoms, nous n'avons connaissance que d'un travail, celui de Frank (2009) qui a simulé des temps de lecture des participants avec un modèle basé sur des connaissances du monde et des contraintes syntaxiques (voir la Section 4.4.2 pour plus d'informations). En plus de ce travail, on peut citer deux travaux qui traitent de la résolution de la coréférence, une problématique liée à celle de la résolution des pronoms : Dubey, Keller et Sturt (2013) et Jaffe, Shain et Schuler (2018).

L'absence de travaux sur la modélisation cognitive et computationnelle des pronoms contraste avec le grand nombre d'études psycholinguistiques sur la résolution de pronoms et l'attention que le sujet a reçu dans la communauté du TAL. Depuis les années 1970, la résolution des pronoms est un sujet populaire dans les deux domaines scientifiques. Nous pensons qu'il est temps maintenant exploiter cette littérature existante pour développer des modèles cognitifs et computationnels de la résolution des pronoms. D'une part, le domaine de la psycholinguistique nous fournit des théories sur la résolution des pronoms et des données cognitives des humains. D'une autre part, le TAL nous donne de nombreux algorithmes qui peuvent être utilisés pour implémenter des modèles. La modélisation cognitive et computationnelle à son tour peut servir à évaluer la plausibilité des théories et être une nouvelle source d'inspiration pour le domaine du TAL.

Nous pensons que les modèles cognitifs et computationnels de la résolution des pronoms peuvent surtout aider à mieux expliquer comment la résolution de pronoms interagit avec la structure du discours. La résolution des pronoms est influencée par de nombreux facteurs. Dans la littérature, souvent on argumente que la structure du discours joue un rôle clé, que c'est le contexte plus large qui guide la résolution des pronoms. Mais quand on regarde de plus près les stimuli utilisés dans les expériences psycholinguistiques, des structures discursives élaborées sont absentes : les stimuli excèdent rarement la longueur d'une phrase. Des études plus centrées sur la structure discursive utilisent parfois des contextes de trois phrases, mais les stimuli restent très courts. Même leur petite taille s'explique par une nécessité d'exclure des facteurs en dehors de ceux étudiés dans les expériences, il est problématique que les théories qui s'appuient sur des principes discursifs se basent

sur des résultats qui viennent d'expériences où le discours est artificiel est très limité. Nous pensons que les modèles cognitifs et computationnels peuvent aider à surmonter ce problème grâce aux manipulations sur corpus.

Cette thèse est organisée en quatre chapitres dans lesquels nous présentons nos expériences. Chaque chapitre fournit une approche différente vis-à-vis de la modélisation cognitive et computationnelle de la résolution des pronoms.

Dans le Chapitre 2, nous avons cherché à savoir comment les corpus de temps de lecture peuvent être utilisés pour l'étude de la résolution des pronoms. Ces corpus constituent des ressources précieuses pour évaluer les modèles cognitifs et computationnels parce qu'ils contiennent à la fois du texte naturel et des mesures de traitement cognitif par les lecteurs humains. Cependant, l'exploitation de ces corpus pour étudier la résolution des pronoms n'est pas simple. Nous avons du apporter des réponses à diverses questions dont : comment peut-on mesurer le temps de lecture de la résolution des pronoms ? Et, quel modèle statistique est adéquat ? Pour l'expérience du Chapitre 2 nous avons utilisé le Dundee Corpus (Kennedy, Hill et Pynte, 2003), un corpus en langue anglaise d'environ 50 000 tokens lus par dix locuteurs natifs dont tous les mouvements oculaires ont été enregistrés. Nous avons annoté tous les pronoms de ce corpus et utilisé les 1 109 pronoms anaphoriques trouvés pour étudier la résolution des pronoms en lecture naturelle. Nous avons regardé si les facteurs révélés lors des expériences contrôlées — appelés biais de résolution — avaient une influence en lecture naturelle. Des exemples de facteurs sont la distance entre le pronom et son référent, ou la fonction grammaticale du référent et du pronom. Dans notre étude, nous avons essayé de retrouver les effets de ces facteurs décrits dans la littérature psycholinguistique. D'une part, le fait de retrouver ces effets démontre leur robustesse, d'autre part, cela confirmerait que nous avions trouvé une méthode adéquate pour explorer la résolution des pronoms dans les corpus de mouvements oculaires.

Il s'est avéré que la modélisation statistique des temps de lecture pour les pronoms est très délicate. Nous avons utilisé des modèles linéaires à effets mixtes. Cette décision était inspirée par d'autres études faites sur le Dundee Corpus présentées dans la littérature. La question principale a été de déterminer à quel endroit apparait l'effet de la résolution des pronoms dans les temps de lecture. Pour répondre à cette question, nous avons testé les temps de lecture de divers mots autour des pronoms. Nous avons constaté que les effets sont souvent tardifs et ne se manifestent pas souvent dans les temps de lecture des pronoms eux-mêmes. Toutefois, il est difficile de déterminer de combien les effets sont retardés. Il semble que différents facteurs aient des délais différents.

Un problème majeur de notre étude est que nous ne pouvons pas dire avec certitude si les résultats que nous avons trouvés sont statistiquement significatifs. La pratique de tester plusieurs mots autour du pronom vient avec un risque augmenté d'erreurs statistiques. Pour réduire ce risque, il serait nécessaire de tester à un niveau de significativité inférieur à 5%. Cependant, dans notre étude, aucun effet ne perdure après cette correction statistique. Les leçons que nous tirons dans le Chapitre 2 sont qu'il est plausible que les effets connus depuis la littérature psycholinguistique se manifestent également en lecture naturelle mais les effets sont souvent non locaux par rapport au pronom. Pour cette raison il est donc nécessaire de réfléchir à nouveau si les temps de lecture classiques sont la meilleure façon de mesurer la charge cognitive liée à la résolution des pronoms et explorer des nouvelles méthodes statistiques qui n'aient pas besoin de corrections sévères.

Dans le Chapitre 3 nous regardons si les programmes du TAL peuvent constituer des modèles cognitivement plausibles pour la résolution des pronoms. Nous avons cherché à savoir si les biais de résolution révélés lors des expériences psycholinguistiques peuvent être simulés avec un algorithme d'apprentissage machine entraîné sur un corpus. Nous avons comparé les résultats d'humains sur une tâche de résolution de pronoms ambigus à ceux de notre programme. Lors de cette expérience, nous nous sommes concentrée sur deux biais de résolution pour lesquels deux études psycholinguistiques ont rapporté des résultats. Notre modèle computationnel a été capable de simuler les résultats pour les deux biais dans les deux études psycholinguistiques. Nous concluons que les biais appris sur corpus sont comparables aux biais de résolution des humains.

Dans le Chapitre 4 nous avons étudié une métrique de coût — une mesure de charge cognitive — basée sur la Théorie de l'Information (Shannon et Weaver, 1949). Il a été montré que les métriques de coût comme la surprise syntaxique (Hale, 2001) sont des prédicteurs significatifs de temps de lecture. Dans le Chapitre 4, nous avons cherché comment utiliser la Théorie de l'Information pour prédire le coût cognitif dû à la résolution des pronoms. Nous émettons l'hypothèse que la charge cognitive liée à la résolution des pronoms est déterminée par la compétition qui existe entre antécédents potentiels d'un pronom. Nous soutenons que cette compétition peut être mesurée en utilisant la notion de l'entropie, connue pour être une mesure de l'ambiguïté. Nous proposons de mesurer l'entropie des pronoms en utilisant des sytèmes de résolution probabilistes développés dans la communauté du TAL (Soon, Ng et Lim, 2001 ; Lee et al., 2017). Nous avons évalué notre métrique de coût sur les pronoms anaphoriques du Dundee Corpus. Afin d'éviter des problèmes dus au fait que la résolution des pronoms ne sont pas un processus local et des problèmes liés aux méthodes statistiques utilisées en Chapitre 2, nous avons utilisé d'autres mesures que les temps de lecture. Au lieu de prédire les temps de lecture, nous avons choisi de prédire si les participants fixent les pronoms. Cela nous donne plus de points de données et un deuxième avantage est que cela rend la mesure de résolution de pronom plus locale : nous n'avons pas besoin de mesurer les mots autour du pronom. Nous avons aussi changé de méthode statistique par rapport au Chapitre 2. Au lieu d'utiliser un modèle linéaire à effets mixtes dans un cadre de statistiques fréquentiste, nous avons choisi d'utiliser des modèles mixtes généralisés dans un cadre de statistiques Bayesiennes. Nos résultats ont montré que notre métrique de coût basée sur l'entropie était en effet un facteur d'influence sur le comportement de lecture des pronoms. Cela appuie notre hypothèse qu'une plus grande compétition parmi les antécédents potentiels d'un pronom, augmente la charge cognitive de celui-ci. Nous concluons que les métriques de coût basées sur la Théorie de l'Information sont aussi pertinentes pour la résolution des pronoms et que la compétition des potentiels antécédents doit être prise en compte par les modèles cognitifs de la résolution des pronoms.

Dans le Chapitre 5 nous faisons les premiers pas pour utiliser les modèles cognitifs et computationnels comme moyen d'évaluation des théories sur la résolution des pronoms. Nous avons choisi d'implémenter une théorie récente sur l'interprétation des pronoms émise par Kehler et Rohde (2013). Nous avons choisi cette théorie parce qu'elle tient compte de multiple facteurs d'influence dans le processus de la résolution des pronoms qui sont tous combinés dans une seule formule mathématique. Plus précisément, la résolution des pronoms est modélisée comme une probabilité conditionnelle (choix d'un antécédent étant donné un pronom). Le théorème de

Bayes permet de décomposer cette probabilité en deux autres. Selon Kehler et Rohde les deux probabilités correspondent à deux types de facteurs d'influence différents : la saillance et la connaissance du monde. Les preuves que Kehler et Rohde citent à l'appui de leur théorie viennent tous d'expériences de complétion : des études dans lesquelles les participants complètent un questionnaire.

Nous sommes d'avis qu'une évaluation sur corpus de cette théorie est importante, afin de vérifier sa robustesse. Dans l'expérience présentée en Chapitre 5 nous avons vérifié la version faible de la théorie. C'est-à-dire que nous avons étudié si la résolution des pronoms peut être modélisée en tant que la probabilité conditionnelle d'un antécédent étant donné un pronom et pas si la résolution des pronoms est en effet influencée par la saillance et les connaissances du monde, comme proposé par la version forte de la théorie. Nous avons utilisé le Provo Corpus (Luke et Christianson 2016) pour notre expérience. Ce corpus a la particularité de contenir des temps de lecture, ainsi que des données d'une *cloze task*, un jeu de devinette où des participants humains lisent le début d'un texte et doivent deviner quel mot suivra. Ces données de *cloze task* nous ont servi à obtenir une estimation grossière des paramètres de la formule mathématique de la théorie. Les données de temps de lecture ont ensuite servi à évaluer cette formule. Nos résultats sont en faveur de l'hypothèse que la résolution des pronoms peut être formulée en tant que probabilité conditionnelle. Cependant, l'estimation de nos paramètres était très bruitée, ce qui nous a emmené à conclure qu'il vaut la peine d'examiner la théorie de Kehler et Rohde plus en profondeur d'une manière plus sophistiquée et coûteuse. Dans les perspectives du chapitre nous proposons des solutions afin d'arriver à une évaluation plus précise et nous envisageons la façon dont la version forte de la théorie — qui précise que la résolution des pronoms est influencée par la saillance et les connaissances du monde — pourrait être évaluée grâce aux modèles cognitifs et computationnels.

En somme, dans notre thèse nous avons présenté différents modèles cognitifs et computationnels de la résolution des pronoms. L'absence de littérature sur ce sujet nous a amené à faire beaucoup de travail exploratoire. Cela nous a donné une meilleure compréhension comment les corpus de mouvements oculaires peuvent être utilisés pour étudier la résolution des pronoms. Nos expériences ont démontré que les systèmes développés pour le traitement automatique de langue peuvent servir de modèles cognitifs computationnels et que la Théorie de l'Information est pertinente pour expliquer la charge cognitive des pronoms. Nous faisons l'hypothèse que ces résultats peuvent être expliqués par le fait que la résolution des pronoms par les humains soit un processus qui est sensible aux phénomènes statistiques du langage.

Acknowledgements

First of all, I want to thank Saveria Colonna, Vera Demberg, Laurent Prévot and Aline Villavicencio to have accepted to be in the jury. Special thanks go to Aline Villavicencio and Laurent Prévot for their reviews. I am also grateful that Laurent accepted the role of president of the jury.

I wish to thank my thesis supervisor, Pascal Amsili, for his guidance from the beginning to the end of this thesis. He dedicated a lot of his time to our long discussions and his advice was of great value for this thesis but he was also an excellent tutor in teaching and other academic matters. I want to thank him for the freedom he gave me in my research, for the solutions he provided for problems, for his great capacity to listen and for his constructive criticism. I also thank my co-supervisor, Adeline Nazarenko, for the trust she had in the project and the warm welcome she gave me in the Laboratoire d'Informatique de Paris Nord during my first year of the PhD.

I thank the members of my thesis monitoring committee, Saveria Colonna and Isabelle Tellier, for their yearly evaluation and their critiques. I am very sad that Isabelle passed away suddenly and could not see the end result but I keep a beautiful memory of her.

I thank my colleagues at the Laboratoire de Linguistique Formelle. In particular, I wish to thank Barbara Hemforth, Benoît Crabbé and Olivier Bonami for the interesting discussions we had on my thesis topic. Special thanks go to Yair Haendler for teaching me Bayesian statistics and helping me with the statistical models in the thesis. Many thanks also to Amandine Martinez, for her help with the annotation of the Dundee Corpus. Among my colleagues, my fellow PhD-student take a very important place. Special thanks go to Maximin Coavoux for his friendship and his corrections and to Sacha Beniamine, Timothée Bernard, Vincent Ségonne, Céline Pozniak, Loïc Grobol and Charlotte Hauser for their friendship and companionship. Thanks also to all my others colleagues, in particular to Alexandre Roulois and the team of engineers for the technical support and the administrative staff for helping organising the thesis defence. I also wish to thank my former teachers from the master's program Linguistique Informatique, Laurence Danlos and Marie Candito, to have inspired me during my Master's.

I wish to thank in particular Vera Demberg, her research group and the SFD 1102 for the very warm welcome they gave me during my research stay at the Universität des Saarlandes financed by the Labex Empirical Foundations of Linguistics. The research stay was very instructive thanks to Vera's and her colleagues' advice and the many presentations and seminars that were close to my thesis topic. Thanks to Dave, Frances, Wei, Tony, Ekaterina, Katja, Merel, Gabi, Elli, Marc, Sébastien, Michael and all the other kind people I met there.

I also thank my doctoral school, Frontières du Vivant, for providing an extensive interdisciplinary doctoral education, as well as additional travel funding that allowed me to present different parts of this thesis at French and international conferences. Thanks to all my fellow FdV-students for their friendship and the great time we shared at the CRI.

And last but not least, thanks to all my other friends and family. You were of great support. Thank you for always believing in me and sharing this happy moment. Thank you Daphne, Sophia, Anna, Gabrielle, Živa, Chantal and Emmanuele! Thank you Florian, Ivolien, Jan, Vincent, Leo, Martine, Pascal and Ernest!

This thesis was financed by the grant Double Culture of the Université Sorbonne Paris Cité and was also supported by funds of the *Laboratoire d'excellence* Empirical Foundations of Linguistics. It is written thanks to the LaTeX template Master/-Doctoral thesis created by Steve Gunn and Sunil Patel shared under the licence of Creative Commons 3.0.

Contents

Chapter 1

Introduction

This thesis is about cognitive computational models: models that have the objective to simulate human language processing. Cognitive computational modelling is an interdisciplinary field influenced by psycholinguistics, computational linguistics and cognitive sciences. This thesis is therefore related to all these domains.

In this thesis, we focus on models for pronoun resolution: the process of finding the antecedent of an anaphoric pronoun. Whereas pronoun resolution is not at all a new topic — countless studies have been done — it has almost never been investigated by the means of cognitive computational models. We believe that cognitive computational models have a big potential to boost research on pronoun resolution. Therefore, the goal of our thesis is to develop such models.

We explored the use of various corpora, different statistical methods and multiple algorithms. This exploration has lead us to develop various models, but above all, it helped us to gain a better understanding of which role cognitive computational models could play in the study of pronoun resolution.

In the following sections we discuss first what cognitive computational models are and what they can add to the field of psycholinguistics and Natural Language Processing. Then, we discuss what it would mean specifically for the topic of pronoun resolution to be investigated by means of cognitive computational models. We finish the introduction with a brief overview of the contributions of this thesis.

1.1 Cognitive Computational Models

The goal of cognitive computational models is to simulate human behaviour with computer programs. A program should give a response that comes as close as possible to a human response when it receives stimuli as input. The computer program calculates its response based on a modelisation of the phenomena present in the stimuli. By confronting the program's response to the human response, the plausibility of the assumptions underlying the modelisation can be evaluated.

It is important that only the *plausibility* of the theory that underlies the modelisation can be evaluated in this way. Indeed, the fact that the program's responses and the human responses are similar does not mean that the program models the process accurately. Similarly, a failure to simulate human behaviour does not imply that the theoretical assumptions on which the modelisation is based are wrong. It could simply be the case that the computer program lacks accuracy and therefore fails in spite of correct theoretical assumptions.

Despite the fact that cognitive computational models are not a means to obtain hard proof for a theory, they have their usefulness. First, they provide theories with

implementations. It seems trivial, but it is not. When a theory is implemented, every prediction a theory makes must be translated into actions of the computer program. This implies that all details of the theory must be thought through because the program must know at every point what it must do exactly.

A second application of cognitive computational modelling is the comparison of theories. When two theories are implemented, we can decide which one better simulates human behaviour. When the implementation of theory A leads to a better fit with the data than the implementation of theory B, theory A is assumed to be more plausible than theory B. This way of comparing theories is widely used in the field of cognitive computational modelling.

Evaluating cognitive computational models can go hand in hand with testing on corpus data. Most theories of language processing support their claims with evidence from psycholinguistic studies. The stimuli in these studies are designed in such a way that other linguistic phenomena cannot be of influence. But in a natural setting, many phenomena play a role at the same time and interact with each other. A corpus — on condition that it has not been specially designed to include or exclude linguistic phenomena — allows to study how well a theory can predict language processing in a natural setting. In other words, it is a proof of robustness: if the theory can still make correct predictions despite of other linguistic phenomena playing a role, it is robust. Cognitive computational models can therefore play an important role in testing the robustness of theories.

A last reason why cognitive computational modelling is useful is that it can inspire solutions in artificial intelligence. Artificial intelligence for linguistics is called Natural Language Processing (NLP). The goal of artificial intelligence is to develop computer programs that are able to do human tasks as well as humans can — or even better. NLP-programs are developed for tasks such as machine translation, grammatical analysis of sentences and automatic reasoning. If a better modelisation for human processing can be reached by implementing cognitive computational models, it can be a great source of inspiration for the NLP-field. Especially because it provides implementations of theories together with evaluations of the plausibility and the robustness of these theories.

1.2 Pronoun Resolution

Pronoun resolution is an anaphora resolution process. According to Van Deemter and Kibble (2000), "*an NP α_1 is said to take an NP α_2 as its anaphoric antecedent if and only if α_1 depends on α_2 for its interpretation*".[1] The case in which a pronoun refers back to an antecedent earlier mentioned in the text is thus a type of anaphora resolution. A pronoun carries very little semantic information by itself and obtains its interpretation by the resolution process. The key question in all studies on pronoun resolution is therefore: what are the mechanisms behind the process of pronoun resolution?

In modern works from the field of NLP, anaphora resolution and coreference resolution are often lumped together. However, technically speaking, they are not exactly the same phenomenon. Van Deemter and Kibble (2000) state that "*α_1 and α_2 corefer if and only if Referent(α_1) = Referent(α_2)*". Whereas the relation of coreference is thus reciprocal, it is asymmetrical for anaphora relation. Moreover, the anaphoric

[1] NP stands for Noun Phrase.

relation is always subject to a resolution process, while the two referents involved in a coreference relation do not always need this resolution process: it is sufficient that they refer to the same extra-linguistic entity.

In this thesis, we made the choice to focus only on pronoun resolution because the anaphoric relation between the pronoun and its antecedent is clearly specified, in contrast to coreference resolution that includes multiple phenomena. In addition, the study of pronoun resolution is well anchored in the psycholinguistic literature while coreference resolution is not.

Cognitive computational models for pronoun resolution are a new topic. Whereas there is literature about cognitive computational modelling for syntactic processing and lexical processing, cognitive computational models for other linguistic subjects are very scarce and the literature about cognitive computational models of pronoun resolution is (nearly[2]) non-existent. There are a few cognitive computational models on coreference resolution (Dubey, Keller, and Sturt, 2013; Jaffe, Shain, and Schuler, 2018) but the topic has not been investigated very much yet.

The absence of cognitive computational models of pronoun resolution contrasts sharply with the large number of psycholinguistic studies on pronoun resolution and the attention that the topic has received from the NLP-community. Since the 1970's, it has been a popular topic in both these communities. We think that the time has now come to take advantage of this existing literature to develop cognitive computational models of pronoun resolution. On the one hand, the field of psycholinguistics provides us with theories about pronoun resolution and data of human processing. On the other hand, NLP provides us with many pronoun resolutions algorithms that can be used to implement models. Cognitive computational models of pronoun resolution can then, in their turn, help to evaluate the plausibility of the processing theories and provide new inspiration for the NLP-community.

We think that cognitive computational models for pronoun resolution can help specifically to explain pronoun resolution on a discourse level. Pronoun resolution is influenced by many features. In the literature, it is often argued that the discourse structure plays a crucial role in the process. This reflects the idea that pronoun resolution depends heavily on the larger discourse context. Yet, when we look at experimental stimuli of psycholinguistic experiments, it is difficult to speak of a discourse structure because the stimuli typically do not exceed the length of one sentence. Some studies that are more careful about discourse might use stimuli with a length of three sentences but the texts stay very short. Even though we can perfectly understand why the texts are short — it has to do with controlling factors outside the scope of study — it remains problematic for theories making claims about discourse structure to back up on evidence that has been gathered from experiments in which discourse is artificial and very limited. We believe that because cognitive computational models can be developed to run on corpus data, they can help to overcome this problem and be used to focus more on discourse structure.

In addition to going towards a discourse account of pronoun resolution, implementing theories would be beneficial to specify them better. Indeed, theories with a discourse account for pronoun resolution, such as *Centering Theory* (Grosz, Weinstein, and Joshi, 1995), or *Accessibility Theory* (Ariel, 1991) contain some abstract concepts, such as *saliency* that are difficult to measure because they are not fully

[2]We know about one cognitive computational model of pronoun resolution: Frank et al. (2009), who simulated reading times of participants using a model based on world knowledge and syntactic constraints. See Section 4.4.2 for more information.

specified. Having cognitive models that actually implement these abstract concepts would help to learn more what is it that makes something *salient*.

In short, we believe it is worth exploring cognitive computational modelling of pronoun resolution because it allows us to enhance current theories of pronoun resolution. Abstract notions of the theories can be investigated better because computational modelling requires an implementation. In addition, the theories' robustness and their ability to scale up to corpus data can be addressed. Furthermore, we can expect a beneficial interaction between the models and the NLP-community: the NLP-community can first provide powerful and robust algorithms that can be used for the cognitive programs and then the results of the cognitive modelling can inspire new models used by the NLP-community.

Of course, all these expected benefits come with their counterparts. We experienced many difficulties specifically related to pronoun resolution. For example: pronoun resolution is influenced by many factors from the fields of syntax, semantics, pragmatics and discourse. It is difficult to account for all these linguistic levels in one model. Another example of a problem we encountered is that the phenomenon of pronoun resolution is not local: resolution does not only take place when a pronoun is encountered. When we studied pronoun resolution using reading data: the resolution process has an influence on how fast the words following the pronoun are read. Moreover, it is possible that encountering the antecedent before the pronoun can influence expectations about the appearance of a pronoun later in the text and therefore influence pronoun resolution in advance. The experiments presented in this thesis made clear what the challenges are and provide solutions to some of these problems.

1.3 Contributions of the Thesis

This thesis is organised around four main chapters in which our experiments are presented. Each chapter approaches cognitive computational models for pronoun resolution in its own way. In this section, we briefly discuss the contributions of each one of them.

In Chapter 2, we explore how reading time corpora can be used to study pronoun resolution. Reading time corpora are an interesting source to evaluate cognitive computational models on because they contain natural texts and measurements of cognitive processing — reading data — at the same time. However, using this type of data to explore the subject of pronoun resolution is not trivial. Questions we needed to answer were for example: how do we measure reading times for pronoun resolution? And, what type of statistical model is appropriate? For this study, we used the Dundee Corpus (Kennedy, Hill, and Pynte, 2003), a corpus in English of about 50 000 tokens that is read by ten native English speakers of whom all eye-movements have been recorded. We annotated all the pronouns of this corpus and used the 1 109 anaphoric pronouns of the corpus to study pronoun resolution in natural text reading. We looked whether linguistic features of the text that were shown to be of influence in psycholinguistic studies — called resolution biases — were also of influence in natural text reading. Examples of such linguistic features were the distance between the pronoun and its antecedent and the grammatical function of the referent and the pronoun. In the study we tried to find back the effects described in the psycholinguistic literature in the data of the Dundee Corpus. On the one hand,

finding these effects in natural data would confirm the robustness of the effects. On the other hand, it would show that we found a method to study pronoun resolution in eye-tracking corpora.

It turned out that the statistical modelling of the reading time data of pronoun resolution was very challenging. We used linear mixed modelling, inspired by previous studies on the Dundee Corpus. One of the most important challenges we faced was to determine where the effect of pronoun resolution shows. To find it out, we tested the reading times of multiple words around the pronoun. We found that the effects are often delayed, not showing in the reading times of the pronoun itself — the pronoun is often not fixated at all — but later in the text. However, it stays difficult to determine how much later exactly the effect occurs: different factors of influence seem to have their own delay.

An important issue with our study is that we cannot grant that the effects we found are statistically significant. Testing multiple words around the pronoun comes with an inflated risk of statistical errors. To reduce the risk, it would be necessary to test at a lower level of significance than 5%. But in our study, no effect survives this statistical correction. The lessons we learn from the study in Chapter 2 are: it is plausible that effects known from psycholinguistic literature are also of influence in natural text reading but the effects are often non-local to the pronoun. Therefore, we have to reconsider whether classical reading times are the best way to measure the difficulty of pronoun resolution and exploit new statistical methods to prevent too severe statistical corrections.

Chapter 3 investigates whether NLP-programs can serve as cognitively plausible models for pronoun resolution. In a search for computational models of pronoun resolution, the NLP-literature is a good starting point, but the question remains whether the systems from NLP are cognitively plausible. We investigated whether resolution biases observed in psycholinguistic experiments can be simulated correctly with a machine learning algorithm trained on corpus data. We compared the results of humans on a task of ambiguous pronoun resolution to those of the model. We focussed on two resolution biases for which two psycholinguistic studies reported results. We show that our computational model is able to simulate the results of both studies accurately. We conclude that biases learned on corpus data are comparable to pronoun resolution biases that humans experience and that a computer model implementation can help to specify theoretical claims.

In Chapter 4, we investigate a cost metric — a measure of cognitive load — based on Information Theory (Shannon and Weaver, 1949). Cost metrics, such as syntactic surprisal (Hale, 2001) have been shown significant predictors of reading times. We reflect on how Information Theory can be relevant to measure processing cost of pronoun resolution. We put forward a hypothesis that states that the cost from pronoun resolution is determined by the competition amongst the antecedent candidates of a pronoun. We argue that this can be measured using the notion of entropy: also known as a measure of ambiguity. We measure the entropy of pronoun resolution thanks to coreference resolution systems from the NLP-community (Soon, Ng, and Lim, 2001; Lee et al., 2017). Then, we test our hypothesis on the pronouns we annotated in the Dundee Corpus. To prevent the problems with the non-local reading times of pronoun resolution and the problems with the statistical models we faced in Chapter 2, we used another measure of reading behaviour than regular reading times. Instead of predicting reading times, we predicted whether participants fixated the pronouns in the corpus. This leads to more data points, and as an additional

benefit, the data remains local: we only have to measure the pronoun and not the words around it. We also shifted to another type of statistical framework. Instead of using linear mixed effects models in a frequentist framework, we opted for generalized linear mixed effect models in a Bayesian statistics framework. This allowed us to test a more sophisticated random effect structure and avoid statistical correction. We found that our entropy cost metric was indeed a factor of influence, confirming that more competition amongst antecedent candidates increases the probability that a participant fixates the pronoun. We concluded that information theoretical cost metrics are relevant for pronoun resolution and that the competition amongst antecedent candidates is a factor that has to be taken into account in cognitive models of pronoun resolution.

In Chapter 5, we present work in which we move further towards testing theories. We make a start with testing an existing theory from the psycholinguistic literature. We choose a recent theory of pronoun interpretation formulated by Kehler and Rohde (2013). Our motivation to choose this model was that it takes into account multiple factors of influence on pronoun resolution and that these factors are combined together in one mathematical formula. More precisely, pronoun resolution is modelled as the conditional probability of choosing a referent given that a pronoun is encountered. Bayes' Theorem then allows to decompose this probability into two new probabilities. Kehler and Rohde associate both of the probabilities with a type of pronoun resolution factors: *saliency* and *world knowledge*. The evidence that Kehler and Rohde report for their theory comes from completion experiments: participants have to complete sentences in a questionnaire.

We believe that a corpus evaluation is important to evaluate this theory's robustness. In the experiment proposed in Chapter 5, we make a first step towards such an evaluation. We evaluated the weak version of the model. That is to say, we investigated whether pronoun resolution could be modelled as the conditional probability of a referent given the pronoun and not whether pronoun resolution is influenced by both saliency and world knowledge. We used the the Provo Corpus (Luke and Christianson, 2016) in our study. This corpus has the particularity to contain both reading time measurements from eye-tracking and cloze task data[3]. We used the cloze task data to obtain a rough estimation of the parameters of the model and we evaluated the model on the reading times. We found some results in favour of the conditional probability assumption of the model but we also saw that the estimation of our parameters was very noisy. It lead us to conclude that it is worthwhile to examine Kehler and Rohde's theory in a more fine grained manner. In the perspective section of the chapter, we discuss how we can obtain more precise estimations of the parameters of the model and how we can also evaluate the claim that the two probabilities of the model are indeed influenced by saliency and world knowledge.

[3] A cloze task is a guessing game: participants are presented with the first word of a text and then have to guess the next word. Once they gave their answer, they are presented with the actual word and have to guess the next word. This goes on until the end of the text. Cloze task data for a corpus are thus a collection of human guesses of each word in a text.

Chapter 2

From Experimental to Corpus Data

2.1 Introduction

In this chapter, we present a study that aims at using uncontrolled corpus data enriched with cognitive measurements — for example reading time — to study pronoun resolution. We argue that the study of this type of data is an interesting complement to psycholinguistic studies.

In the field of psycholinguistics, there has been a long history of studying anaphora resolution. Psycholinguistic studies on the resolution of ambiguous pronouns have revealed many linguistic configurations that can be of influence on pronoun resolution: linguistic configurations can trigger participants to prefer one resolution over another, even if both resolutions are possible. As examples we can cite the preference for a resolution to an antecedent in the subject position (e.g. Crawley, Stevenson, and Kleinman, 1990), or an antecedent with the same syntactic function (e.g. Smyth, 1994). These two examples are only a fraction of the landscape of influences that have been discussed in the psycholinguistic literature.

In addition to asking people what the antecedent of an ambiguous pronoun is, there are many other methods used in psycholinguistic experiments to study biases on pronoun resolution, e.g. one can measure the time that it takes participants to resolve a non-ambiguous pronoun. Depending on the absence or the presence of the linguistic feature that makes resolution harder or easier, the reaction time can be faster or slower.

Experimental studies are typically conducted in a controlled manner. To study factors of influence on pronoun resolution, researchers create experimental items. To put it simply: items are created in a manner that there is a minimal difference between the items with the factor and the items without the factor. This ensures that if the researchers find a difference between the two types of items, they can quite surely attribute it to the factor they were interested in. Whereas this method has lead to many interesting findings about pronoun resolution that we will discuss in the next section of this chapter, we argue that the downside of this method is that it makes the context in which the pronouns are presented rather artificial.

Pronoun resolution and coreference resolution take place on a discourse level. Even though some pronouns are constrained by syntax, the resolution of most pronouns also demands the integration of context information. We think that the context of the sentence is not sufficient and that a discourse context is necessary. The items used in controlled studies are often very short. Most of the time they consist of one sentence in which the pronoun is presented. In very rare cases, a discourse is constructed containing three sentences. The reason for short items is obviously that

the items need to be controlled: they need to be minimally different except for the tested factor. The longer they are, the more difficult this becomes.

Using corpus data to study pronoun resolution directly addresses these limitations of experimental studies. The pronouns are presented in a discourse context that is not artificially reduced: this matters directly for the study of the influence of distance between the pronoun and the antecedent. But it also matters for other aspects. In experimental items, the number of referential expressions is kept low and the number of expressions compatible with the pronoun is also kept low. The influence of the number of referential expressions could be studied more thoroughly on natural corpus data.

Studying pronoun resolution in natural text can also lead to more insights on the interaction of factors playing a role in pronoun resolution. In a natural context, it is possible that multiple factors play a role at the same time. The study of pronoun resolution on corpus gives insights into the interactions between factors and the intensity of the influence of each factor. Indeed, the influence of factors is difficult to measure from psycholinguistic experiments because efforts have been made to exclude influence of phenomena that are not under study. The filtering out of 'noise' could make the effects found for the factor under study higher than it actually is in natural data.

A last reason to introduce the study of corpus data as a complement of controlled experiments, is about the testing of theories. Many theories about pronoun resolution, such as Centering Theory (Grosz, Joshi, and Weinstein, 1983; Grosz, Weinstein, and Joshi, 1995), or Accessibility Theory (Ariel, 1991) explain pronoun resolution on a discourse level. In order to evaluate these theories better, we think it is necessary to make a step towards the analysis of more natural discourse data.

In summary, there are many reasons why using uncontrolled corpus data is interesting for studying pronoun resolution. However, the goal of this chapter is not to take up all these challenges at once. A very important question that needs to be answered is whether it is feasible to study pronoun resolution on uncontrolled data. To our knowledge, this type of study is new.[1] Because this is the first study of this kind, a clear methodology is not yet available. Therefore, this chapter also focusses on the development of suitable methods. The goal of this chapter is thus to present a first exploration of a corpus of English containing reading data to study pronoun resolution. The experiments we present should be seen as a study of the feasibility.

The reason why we underline that the results in this chapter must be seen as an exploration, is that the study of pronouns on uncontrolled data implies a lot of obstacles — and we faced many. The most obvious obstacle is that the pronouns are uncontrolled: they do not all occur in the same location in the sentence, in the same syntactic structure and their antecedents have varying lexical frequencies. We first need to explore how we can use eye-tracking corpora, despite the noise coming from uncontrolled factors.

We used the Dundee Corpus (Kennedy, Hill, and Pynte, 2003), the largest corpus for English with reading times. The corpus contains 65 small texts that participants read. They only had to read the texts, while their eye-movements were recorded. These texts naturally contained anaphoric pronouns, which are the object of our

[1] Studies about coreference resolution (Dubey, Keller, and Sturt, 2013; Jaffe, Shain, and Schuler, 2018) also used corpus data enriched with reading times, but the approach remains quite different from ours, because we typically focus on pronouns whereas these studies focus on the influence of all forms of coreference.

study. We annotated them, to separate them from non-anaphoric pronouns and to identify their antecedent.

As a start of using eye-tracking corpora to study human pronoun resolution, we tested some of the most well-known resolution biases found for English. As these biases have been found to play a role in multiple studies using different methodologies, the effects are probably robust. Therefore, we can use them to investigate how to exploit eye-tracking corpora: if the effects show in uncontrolled data, we know that the method to explore eye-tracking data we used can overcome noise and that we can thus use it to test the influence of other effects as well.

The chapter first focuses on different resolution biases that have been described in psycholinguistic data in order to get a clear picture about what type of linguistic phenomena influence pronoun resolution (see Section 2.2). Second, we discuss the statistical method we developed to study pronouns on eye-tracking from corpus reading data and the results we obtained (Section 2.3). Third, we describe the Dundee Eye-tracking corpus and the annotation of pronoun resolution we provided (Section 2.4). We close the chapter with a discussion about what we can learn from these experiments and how this research should be continued (Section 2.5.3).

2.2 Pronoun Resolution Biases

In this section, we discuss the preferences, or biases, that play a role in pronoun interpretation and that we decided to use for our model. A preference is defined as a linguistic feature that strongly biases the pronoun interpretation towards one of the possible antecedent candidates of the ambiguous sentence. We use the word preference and bias interchangeably.

The list of preferences does not aim to present all biases discovered to be of influence on pronoun resolution. We give a broad picture about the research that has been done in the field of psycholinguistics. We choose to only include specific biases in our overview that focuses on the English language, because the corpus we investigated in our experiment is in English. Furthermore, this means that we do not discuss how prosody and other phonological phenomena can influence pronoun resolution, nor how semantic features, such as implicit causality or discourse relations can. Whereas there is a very large and interesting literature about these type of factors, it would be outside the scope of this chapter. For more information about the influence of implicit causality and coherence factors, we refer the reader to Section 4.4.2 and Section 5.2.

2.2.1 Bias for the Subject

The bias for a resolution of the pronoun to the syntactic subject has been described very early (Broadbent, 1970; Hobbs, 1976; Clancy, 1980; Frederiksen, 1981). In English, pronouns have more often antecedents in the subject position than in other syntactic positions. This explains the facilitation in pronoun resolution when the antecedent of the pronoun is indeed a syntactic subject. Various experimental studies find an effect of the subject preference (e .g. Crawley, Stevenson, and Kleinman, 1990; Järvikivi et al., 2005). For example, Crawley, Stevenson, and Kleinman (1990)

found that in ambiguous sentences, such as (1) in which the pronoun was not a subject itself, participants preferred to choose the subject when they asked what would be the most likely antecedent.

(1) Richard and Jim were playing cops and robbers on the old playing fields. Their classmate Caroline passed by on her way to the shops. Richard chased Jim round the corner and Caroline ignored him$_{resolve}$.

However, the bias for the subject is questioned by cross-linguistics results. Hemforth et al. (2010) showed that in French, we should rather speak of a bias for the syntactic object. They conducted two experiments: one visual world paradigm experiment and one off-line completion (questionnaire) experiment to compare English, German and French. They used experimental items as in (2), in which the same sentence was translated into the three languages under study. They found that the French participants showed a preference for the antecedent candidate in the object position, whereas the English and the German showed a bias toward the antecedent in the subject position.

(2) a. The postman met the streetsweeper before he went home.
b. Der Briefträger hat den Strassenfeger getroffen bevor er nach Hause ging.
c. Le facteur a rencontré le balayeur avant qu'il rentre à la maison.

Hemforth et al. (2010) explain their findings by the very frequent infinitive construction in French that points unambiguously to the subject (3-a). A similar construction exists in English (3-b) but it is less frequently used. A second important explanation is that in general, pronouns have an antecedent in the syntactic object position in French.

(3) a. Le balayeur a rencontreé le facteur avant de rentrer à la maison.
b. The street-sweeper met the postman before going home.

The fact that the bias for the subject is not universal could indicate that it might actually be an epiphenomenon of another mechanism that operates in pronoun resolution. Often it is argued that the subject is preferred as an antecedent, because the subject position is very *salient*, at least in English (e.g. Ariel, 1991). According to the *saliency vision*, the preference for the subject is a consequence of syntactic subjects being salient in English.

In this chapter, we will stick with the bias for the antecedent in the subject position, rather than finding a measure of saliency. The reason is that the effect of a preference of the subject has been well documented, whereas for the effect of saliency, some discussion remains about what factors exactly contribute to it. Therefore, it is easier to put into practice an implementation for the subject bias than for a saliency bias.

2.2.2 Bias towards the Antecedent in a Parallel Syntactic Function

The origin of this bias lies in the work of Sheldon (1974). She observed that children make use of the parallel syntactic function to interpret relative clauses. She tested children from about four to five years, asking them to perform with toy animals an illustration of sentences, such as (4). She observed that children were better in interpreting sentence in which the attachment of the relative clause was parallel to

the function of the gap in the relative clause. This observation was called the *Parallel Function Hypothesis*.

(4) a. The dog$_i$ {that gap_i jumps over the pig}$_{rel}$ bumps into the lion.
subject relative clause and *i* is the subject of the relative clause
b. The lion$_i$ {that the horse bumps into gap_i}$_{rel}$ jumps over the giraffe.
subject relative clause and *i* is the object of the relative clause
c. The pig bumps into the horse$_i$ {that gap_i jumps over the giraffe}$_{rel}$.
object relative clause and *i* is the subject of the relative clause
d. The dog stands on the horse$_i$ {that the giraffe jumps over gap_i}$_{rel}$.
object relative clause and *i* is the object of the relative clause

In the domain of pronoun interpretation, the bias for the antecedent in a parallel syntactic function is described — in most psycholinguistic works (e .g. Crawley, Stevenson, and Kleinman, 1990; Smyth, 1994) — as resolving a pronoun to an antecedent that has the same syntactic function as the pronoun. Whether a pronoun has a parallel function is, because of this description, quite dependent on the syntactic framework that is worked in. For example: does the syntactic framework consider that a subject of a passive sentence is parallel to a subject of an active sentence? Smyth (1994) showed that the definition of the bias for parallel structure needs even more reflection. He remarked that it is not only the same syntactic function that matters: he found that the effect was stronger when the whole syntactic structure of the sentence of the pronoun and its antecedent were parallel. For example, adjuncts in the sentence have an influence on the strength of the bias towards a parallel interpretation: when adjuncts are parallel as well, the effect is stronger.

As a last remark on this bias, we note that it is not always easy to distinguish the influence of the bias for resolution to the subject from other biases. For example, if both the pronoun and antecedent are subjects, it is not sure whether the bias is due to the subject position or the parallel function. Studies that also tried to control for the subject bias, report conflicting result concerning the influence of the bias for parallel functions. For example Maratsos (1973) and Smyth (1994) found evidence in favour of the parallel function preference, whereas Rondal et al. (1984) and Crawley, Stevenson, and Kleinman (1990) do not find that the parallel function preference is a factor in pronoun interpretation. We think that these conflicting results are probably the consequence of the fact that the bias for resolution to a parallel antecedent is a bit ill-defined.

2.2.3 Bias towards the First Mentionned Antecedent

Another bias that can easily be confused with the bias of interpereting the pronoun as the syntactic subject has been referred to in litterature as the 'first mention preference'. According to Gernsbacher and Hargreaves (1988), Gernsbacher, Hargreaves, and Beeman (1989), and Gernsbacher (1990), being introduced first in a discourse makes an entity more likely to be referred to later in the text. According to them, when building a mental representation of a sentence, the first mentioned entity is the basis onto which further information is stacked. Therefore, there should be a preference for the first mention in pronoun resolution. In a probe recognition experiment, Gernsbacher and Hargreaves (1988) found that a probe was more easily recognized when it was mentioned first, even if it was not the grammatical subject of the stimulus sentence.

The first mention bias and the subject bias are difficult to distinguish in a language such as English, where grammatical subjects are mostly at the beginning of sentences. Järvikivi et al. (2005) tried to untangle the two preferences by conducting a visual world eye-tracking experiment in Finnish. As Finnish is a free-order language where grammatical role is marked by morphological means, they had stimuli with sentences in either the subject-verb-object order and the object-verb-subject order. The authors found that participants had a preference for the subject and the first mention. However, they did not find a preference for the first-mentioned entity in the object-verb-subject sentences. Järvikivi et al. (2005) concluded that both the subject preference and the first mention bias can play a role at the same time. They argue that *"one-factor models are inadequate, and that pronoun resolution is determined by a delicate interplay of several factors."* A similar study on Finnish investigated the influence of grammatical role and order of mention (Kaiser and Trueswell, 2008). This study only found an effect of the grammatical role on personal pronoun interpretation and no effect of the order of mention. Finally, an eye-tracking experiment of Fukumura and Gompel (2015) in which reading time was studied for British English also pointed to an influence of the grammatical role on pronoun interpretation and an absence of the influence of the order of mention.

2.2.4 Bias towards the Most Recently Mentionned Antecedent

The distance between the pronoun and the antecedent is assumed to play an important role in pronoun resolution. Indeed, theories providing a saliency account for pronoun resolution, such as *Centering Theory* (Grosz, Joshi, and Weinstein, 1983; Grosz, Weinstein, and Joshi, 1995) or *Accessibility Theory* (Ariel, 1988; Ariel, 1991) state that (all things being equal) a shorter distance between the pronoun and a potential antecedent makes the antecedent more *salient*. This means that the antecedent is easier to retrieve from memory.

The influence of distance was studied in various psycholinguistic experiments (e.g. Clark and Sengul, 1979; Ehrlich and Rayner, 1983). Clark and Sengul (1979) and Ehrlich and Rayner (1983) recorded reading time for pronouns of which the antecedent occurred one, two or three sentences back. Pronouns of antecedents that were further away needed more reading time.

The reader may have noticed that the bias for an antecedent that is mentioned first might be in contradiction with the bias for an antecedent that is closer. Studying pronoun resolution on natural corpus data could help resolving this issue by giving a more precise definition of distance and first mention. Questions that could be answered using more context are for example: Is distance measured in sentences, clauses, words, or characters? And is it important for first mentioned antecedents that they are mentioned at the beginning of the text, or at the beginning of the paragraph?

2.2.5 The Role of the Frequency of the Antecedent

The lexical frequency of the antecedent plays a role in anaphoric relations. When a pronoun is processed, the antecedent has to be recovered and its lexical frequency can have an influence on the ease of the process (Van Gompel and Majid, 2004). However, different psycholinguistic studies found different effects of the frequency of the antecedent. Two theories predict an opposite effect of frequency. Van Gompel

and Majid (2004) call the first one the *Full Reaccess Hypothesis*. This theory predicts that the reactivation of the antecedent when the pronoun is read is very similar to normal lexical access of words, where infrequent words cause more processing cost (Shillcock, 1982). Hence, an infrequent antecedent will lead to longer reading times. The second theory that Van Gompel and Majid (2004) mention is the *saliency account* (Pynte and Colonna, 2000). This theory predicts that antecedents with lower lexical frequency are more salient (better marked) and therefore easier to recover, evoking shorter reading times. However, the original work of Pynte and Colonna (2000) was about relative clause attachment. They observed that when people read ambiguous sentences in which a relative clause could be attached to two noun phrases, they showed a preference to attach it to the least frequent antecedent, supporting the saliency account. But as Van Gompel and Majid (2004) note, it could be that relative clause attachment is different from pronoun resolution. In their own study on pronoun resolution, Van Gompel and Majid (2004) found shorter reading times for the word after the pronoun for infrequent antecedents. They interpreted these results as the saliency account also being relevant to pronoun resolution.

The question arises whether the preference for a less frequent antecedent can be found in corpus data. The distance between the pronoun and its antecedent in Van Gompel and Majid (2004)'s study was rather short. Would the saliency effect of infrequent antecedent be conserved over a longer distance and could it show in uncontrolled corpus data? We hope that our study can shed some light on these questions.

2.3 Modelling Reading Times for Anaphoric Pronouns

Data about eye-movements can be used to study linguistic processing because many studies found a link between the movements and moment to moment on-line processing (Rayner, 1998). However, modelling reading times for pronouns is not straightforward and it is even more challenging in uncontrolled data. In order to build a model, it is important to consider how the reading zones must be defined, what reading measure is used, how the model controls noise from the uncontrolled data and what data points must be filtered out. In this section, we will explain why these considerations are important and what the possibilities are for the modelling of pronoun resolution.

2.3.1 Zone of Interest for Anaphoric Pronouns

Anaphoric pronouns are in general very short words. This makes attesting reading time for them difficult, because short words are very often not fixated during reading (Rayner, 1998). However, the fact that pronouns are not often fixated does not necessarily mean they are not read and processed. Kennedy and Pynte (2005) found evidence that unfixated words can be processed — at least on a lexical level — when they occur in the parafoveal vision. The *parafoveal vision* is the area that immediately surrounds the center of the fixation. This center is called the *foveal area*, and it constitutes the 2 degrees center of the vision (Rayner, 1998). The parafoveal area is the area from the foveal area until 5 degrees from the center of vision (Rayner, 1998). The vision in the parafoveal area is far from as good as in the foveal area, but it appears to be good enough to do at least some linguistic processing. Therefore, with

respect to pronouns, it could be the case that processing can be initiated when the word before or after the pronoun is fixated.

Another question is whether anaphoric pronouns are processed the moment they are fixated in foveal, or parafoveal vision, or if the processing takes place later in time and shows as a so-called *spill-over effect*. In their studies on pronoun reading, Ehrlich and Rayner (1983) and Van Gompel and Majid (2004) concluded that the retrieving process of the antecedent is initiated where the pronoun is encoded — a fixation on the pronoun itself, or a fixation very near to the pronoun on an adjacent word — but that the processing can be continued later, and show up in the eye-tracking data as a spill-over effect.

To define the region of interest of reading of anaphoric pronouns is — because of all these reasons — very challenging. In the studies presented in this chapter, we will therefore explore a large zone, starting before the pronoun and going until four words after the pronoun.

2.3.2 Reading Times and Metrics for Anaphoric Pronouns

Besides the reading zone, the choice of reading times and reading metrics is also important. Indeed, upon eye-tracking data, many reading times and reading metrics can be defined. We sketch the anatomical mechanisms of reading and explain the reading time metrics by giving a summary of Rayner (1998).

When people read a text, they are making fixations on the text. An individual fixation[2] is on the level of a letter in the text. That is to say, the gaze falls on a letter and not on an entire word. Fixations follow each other very rapidly and every fixation only takes a fraction of a second. The eye 'jumps' quickly from fixation to fixation. These movements are called saccades. On average, the saccade length is about 8 letters. During the saccades, vision is suppressed, meaning that information about the text comes in during the fixations and not during the saccades.

Generally, reading times are calculated on the basis of the fixations. It is assumed that there is a link between the time it takes to read and the difficulty of language processing. Reading times can be calculated by summing up the time of fixations in a given region of interest. The region of interest is the part of text one wants to calculate reading time for. This can be any type of region: a sentence, a word — two types of regions that are often used because the linguistic unit is meaningful — but theoretically, a region could also be two characters.

Reading times are defined as the sum of the duration of certain fixations in the region of interest. There are various ways to determine which fixations to consider. We explain the *first fixation*, the *first pass*, the *regression path* and the *total* reading times, in the following paragraphs and with the help of example (5), Table 2.1 and Table 2.2.

In example (5), we see a sequence of fixations of one participant for the first sentence of the Dundee Corpus (Kennedy, Hill, and Pynte, 2003). The words of the sentence are defined as the reading zones. For every word in the sentence, the fixations that fell on it are written beneath it. Note that saccades do not only go from left to right, but also from right to left. In Table 2.1, for every fixation in the sentence the length in milliseconds is given. Then in Table 2.2 it is explained how

[2]The word 'fixation' is used because readers maintain the gaze for a very small period of time on a letter, but actually there is a constant tremor in the eye, so even during the fixation the eyes are making very small movements.

the fixations of example (5) are used to calculate different reading times and metrics for the region of interest in red. It would be outside the scope of this thesis to give a complete overview of all reading times and metrics. We choose instead to provide a comprehensive introduction for readers that might not be familiar with eye-tracking.

First Fixation

The first fixation reading time is simply the length of the first fixation in the region of interest (Rayner, 1998). If we look at example (5) and we define our region of interest as the word *existence?*[3] we only look at the time of the first fixation on the region. In Table 2.2, we see that fixation number 10 is the first in this region and that it lasted for 173 ms. Therefore, the first fixation reading time is also 173 ms for this region.

First Pass

The first pass reading time is the time spent in the region from entering the region[4] until leaving the region (Rayner, 1998). So, returning to example (5), looking at *existence?*, we see that 10 is the first fixation in this region, followed by 11 and 12 and then we see that the region is left, because fixation 13 falls on the word *enticed*.

Regression Path Duration

The regression path duration is the sum of fixations from the moment the region is entered, until the region is left on the right (the participant continues reading). When we look at example (5), we see that after the fixations 10, 11 and 12 on the region, there is a regression including fixations 13 until 19. Then, the region is fixated again in the fixations 20 and 21, before the next sentence is read. Therefore, the regression path duration is the sum of fixations 10 until 21.

Total Reading Time

The total reading time is simply the sum of the duration of all fixations in the zone. In example (5), fixations 10, 11, 12, 20 and 21 fall in the region of interest, therefore, the durations of all these fixations are summed to obtain total reading time.

Number of Fixations

Instead of looking at the fixation time, the number of fixations could also say something about whether the region of interest is difficult to process or not. An even more crude measure is a boolean metric that says whether the region is fixated or not. These measures can be useful for words that are very short and therefore skipped a lot. When a region is not fixated, the duration of the fixations can not be summed up.[5] Using the number of fixations is a solution to this problem because it allows to include non-fixated reading zones by attributing them a value of 0.

[3] Notice that the question mark is also part of the region. The reason for this choice is that it forms a single visual unit.

[4] Sometimes it is assumed that the region should be entered from the left, but this is not part of the definitions we have come across.

[5] In fact, sometimes a reading time of 0 ms is registered but most researchers treat non-fixated zones as missing values.

(5) Are tourists enticed by these attractions threatening their very existence?
1 2 3,13 4,14,15 5,16,17 6,7,18 9 8,19 10,11,12,20,21

TABLE 2.1: The fixation times for all the fixations of example (5).

Fixation	Word	Fixation Duration (ms)
1	Are	216
2	tourists	156
3	enticed	227
4	these	187
5	attractions	182
6	threatening	96
7	threatening	232
8	very	335
9	their	168
10	existence?	173
11	existence?	188
12	existence?	88
13	enticed	174
14	these	168
15	these	170
16	attractions	271
17	attractions	88
18	threatening	232
19	very	202
20	existence?	222
21	existence?	157

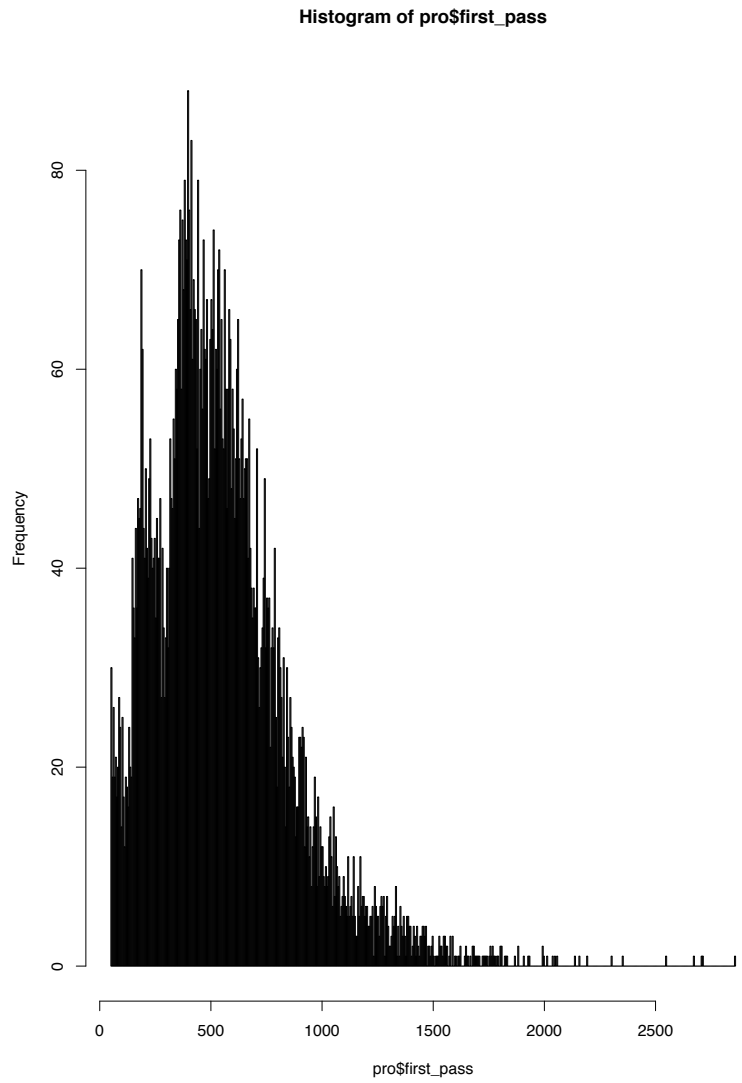
TABLE 2.2: An explanation of some of the most used reading measures, based on example (5), where the word *existence* is the region of interest.

Measure	Definition	Fixations	Outcome
First fixation duration	The duration of the first fixation in the region of interest	10	173 ms
First pass duration	The sum of fixations from the moment of first entering the region until leaving the region	10,11,12	449 ms
Total duration	The sum of all the fixations in the region	10, 11, 12, 20, 21	828 ms
Regression path duration	The sum of all the fixations from the moment of entering the region, until leaving it on the right	10 until 21	2133 ms
Number of fixations	The number of fixations in the region of interest	-	5
Fixated	Boolean variable whether the region was fixation	-	True

Distribution of Reading Times

Another question is to what extent the different reading times and metrics can be explored by statistical models. Traditional reading times, such as the first pass duration, or the total reading time, can be characterized by a skewed positive distribution: all data points are more than 0 and there is a strong concentration of data points with shorter durations. These distributions also contain a tail that is formed by longer durations (see Figure 2.1). The more fixations that are counted in the reading time, the heavier the tail will be. It is important to note that most statistical models assume normal distributions. Because the distribution of reading data is not normal, we must either transform the data or adapt the statistical model. Many adaptations of models are pre-programmed into softwares like R (R Development Core Team, 2008). However, it can happen that in spite of transforming the data or using a preprogrammed distribution, the fit remains poor.

FIGURE 2.1: An illustration of the distribution of reading times: all values are positive and the skewed distribution presents a long tail.



2.3.3 Linear Mixed Effects Models

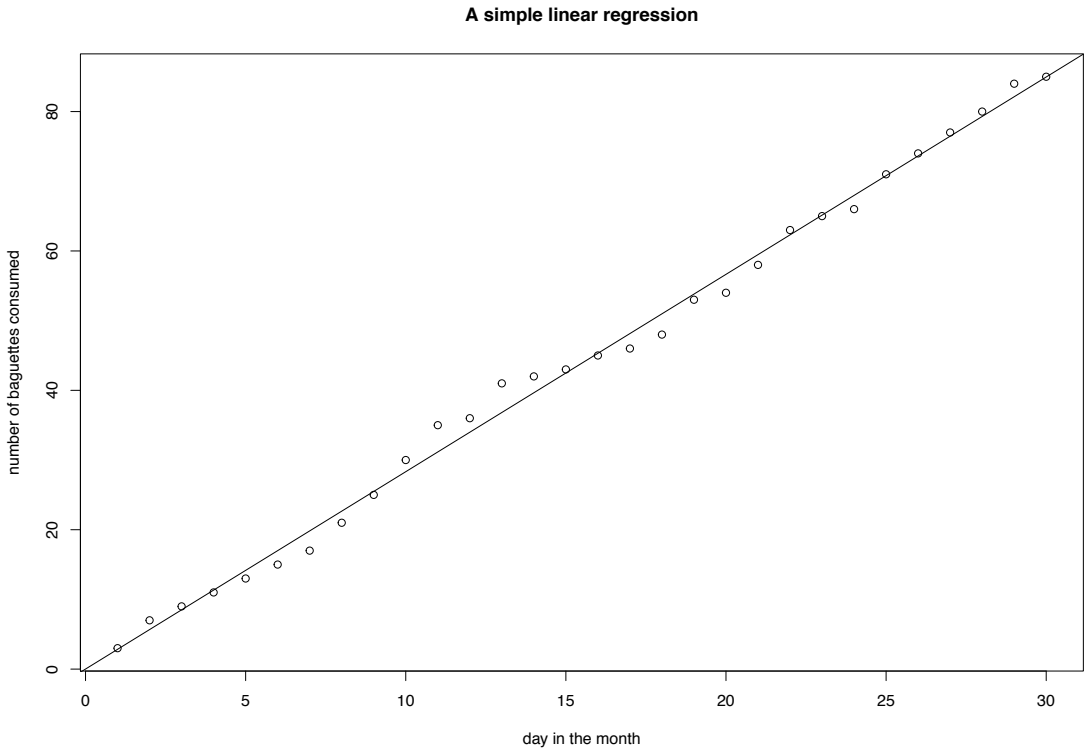
In this chapter, we build Linear Mixed Effects Models. In this subsection, we will explain the properties of such models.

Linear mixed effects models are a special type of linear models. Linear models are better known as linear regression models. The simplest linear regression model is a model with one explanatory variable. This model can be represented by the following formula (Field, 2009):

$$y = \alpha + \beta x \tag{2.1}$$

The parameter α is called the intercept and β the slope, whereas x represents the explanatory variable. An example of such a regression would be the number of

FIGURE 2.2: A toy example of a simple linear regression model. The y-axis represents the number of baguettes consumed in the course of a month. The x-axis represent the day of the month. The straight line is the regression model that has been estimated for the data points.



baguettes a family has been eating during a month – starting at 0 at the beginning of the month,[6] summing the number of baguettes eaten that month. In such a case, the number of the day is the predictor variable. If a family counts each day the number of baguettes they have consumed during the month so far, and obtains the data points of Figure 2.2, a linear model can be estimated using these data points:

$$baguettes = 0 + 2.83 * days \tag{2.2}$$

The relationship between the function and the data points can be explained by the above formula 2.2 plus an error term. The error term is necessary, because not all points lie exactly on the line the model predicted. So, for every data point y_i (the number of baguettes consumed during the month up to day i) we can say that the value is dependent on the intercept α (the number of baguettes consumed when the month starts, that is 0), the number of the day x_i, the slope of the model β that represents an estimation of how many baguettes are consumed in one day, and ϵ_i, the error.

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{2.3}$$

The challenge of building a linear regression model is thus to find the values for α and β. The model that comes as close as possible to the observed data points is called

[6] This example including the data was made-up to illustrate linear regression.

the model with the best *fit*. This model can be found when the error ϵ is minimized with respect to the observed data points. The method of least squares is used for this purpose. For any straight line that can be drawn across the data points, the distance between the line and the points can be measured. Sometimes the line will underestimate a point and sometimes the line will overestimate it. Therefore, there will be positive and negative distances, which are called residuals in the framework of regression. The square values of these residuals are summed (note that the application of the square avoids that the positive and negative distances would cancel out each other). The line with the least sum of squares is the model that it chosen, because it fits the data the best.

The simple linear regression model is easily extended to a multiple regression model including various predictors (Field, 2009):

$$y_i = \alpha + \beta_1 x_{i_1} + \beta_2 x_{i_2} + \cdots + \beta_n x_{i_n} + \epsilon_i \tag{2.4}$$

The parameter α is still the intercept and for every data point y_i, measures for predictor 1, 2, ... n are expressed by $x_{i_1}, x_{i_2}, \ldots x_{i_n}$. The β terms are the coefficients of the predictors. A negative coefficient says that a higher proportion of predictor x leads to a lower outcome variable and for positive coefficients the opposite is true.

To evaluate the quality of a regression model, the model is compared to a baseline model that always predicts the mean of the distribution. The behaviour of the baseline model is: 'no matter what the values of the predictor variables are, the mean is what I predict'. The challenge for the regression model is then to give a better prediction than the mean. To calculate this, the sum of squares of the baseline model is compared to the sum of squares of the regression model. Individual predictors can also be assessed. The question that arises for every β_n-term is: is the coefficient of the regression model different from zero? To answer this question, for every coefficient a t-test is performed. If the t-test is significant, it means that the chance that the predicted is not different from zero is less than 5%.

Normal Distribution of Residuals

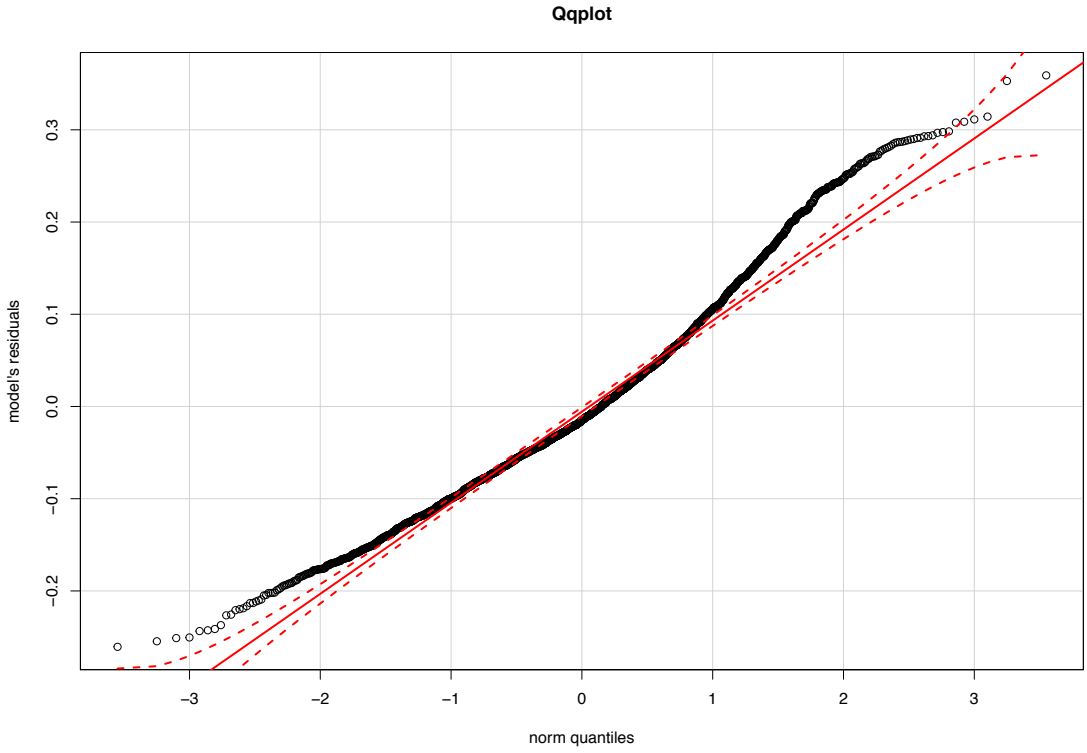
One important thing to look at when using linear modelling, is the distribution of the model's residuals (Field, 2009). Linear regression assumes that the model's residuals are normally distributed. If it is not the case, the conclusions based on the model can be completely erred. A way to see whether the model's residuals follow a normal distribution is to draw a QQ-plot (*quantile-quantile plot*). The quantiles of the data are plotted against the expected quantiles of a normal distribution. See Figure 2.3, for an example of a qq-plot from one of the models in this chapter.

Transformation of Reading Times

The distribution of reading times is not normal, as illustrated in Figure 2.1. Therefore, the distribution of residuals on this data is also unlikely to be normal and therefore reading times without any transformation cannot directly be used as the outcome variable in the linear model. The most common way to obtain a normal distribution of reading times is to apply a log-transformation:

$$\textit{log transformed reading time} = \log(\textit{reading time}) \tag{2.5}$$

FIGURE 2.3: A qq-plot of the residuals of first-pass reading data from the Dundee Corpus. Because the data points follow the straight line, we see that the model's residuals fit the normal distribution quite well.



After the log-transformation, the distribution looks more normal, as is illustrated in Figure 2.4. Besides the log-transformation, there are other possible reading time transformations. For example, the inverse transformation:

$$\textit{inverse transformed reading time} = 1/\textit{reading time} \tag{2.6}$$

The log-transformed is the most widely used transformation and often it is claimed that the inverse transformation might be more suited for self-paced reading times (Rayner, 1998).[7] But it certainly depends on the data. Every researcher needs to check for their experiments which transformation works the best, if they want to use linear modelling.

Recently, software has been developed to deal directly with long tail distributions, such as reading time distributions (Bates et al., 2015). These models are called generalized linear mixed effects models. Different types of distributions can be specified, and the reading time distribution can be appropriately modelled by the Gamma distribution. Because this technique is not used in the experiments presented in this chapter, we will leave further explanation of these generalized mixed effect models to the following chapters of the thesis.

Transformation of Predictor Variables

The predictor variables in regression models can either be numerical, boolean or categorical. Categorical factors are automatically encoded in R with dummy encoding, or contrast encoding. Contrast encoding consists in choosing one category of the values the variable can take as a baseline. Then, all other categories are compared in pairs against this baseline. So for example, if you have the factor *syntactic function* with the values {subject, direct object, indirect object, adjunct}, one of the categories, actually just the first, is taken as the baseline and then all the others are compared pairwise against it: {subject, direct object}, {subject, indirect object}, {subject adjunct}. So there are always the number of categories minus one pairs of comparisons that are made. Note that a comparison like {direct object, indirect object} is not made by this encoding. A solution can be to shuffle the list of categories around, so that another category comes first and serves as a baseline.

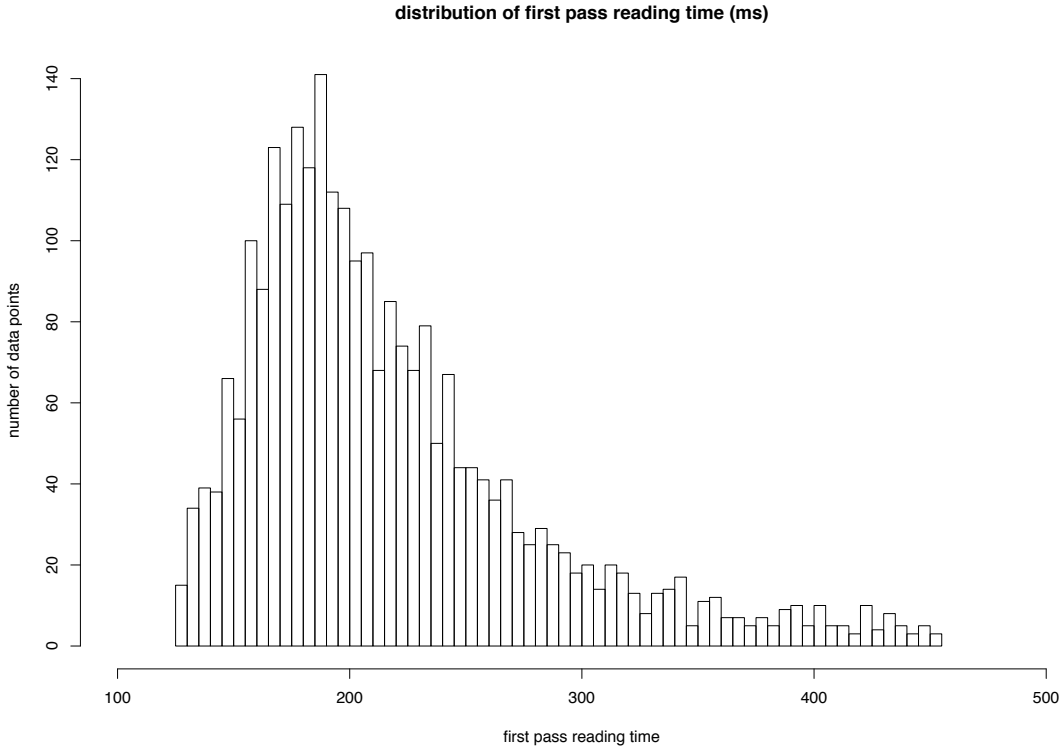
One-hot encoding is an alternative for dummy encoding that encode the influence of each category. It consist in breaking down one categorical variable into multiple boolean variables. So in the example cited above, you would obtain four variables. The one for subject, for example, has the value of 1 when the data point is a subject and 0 otherwise. This is done for all categories. In that way, for each category you obtain a comparison of one versus the rest.

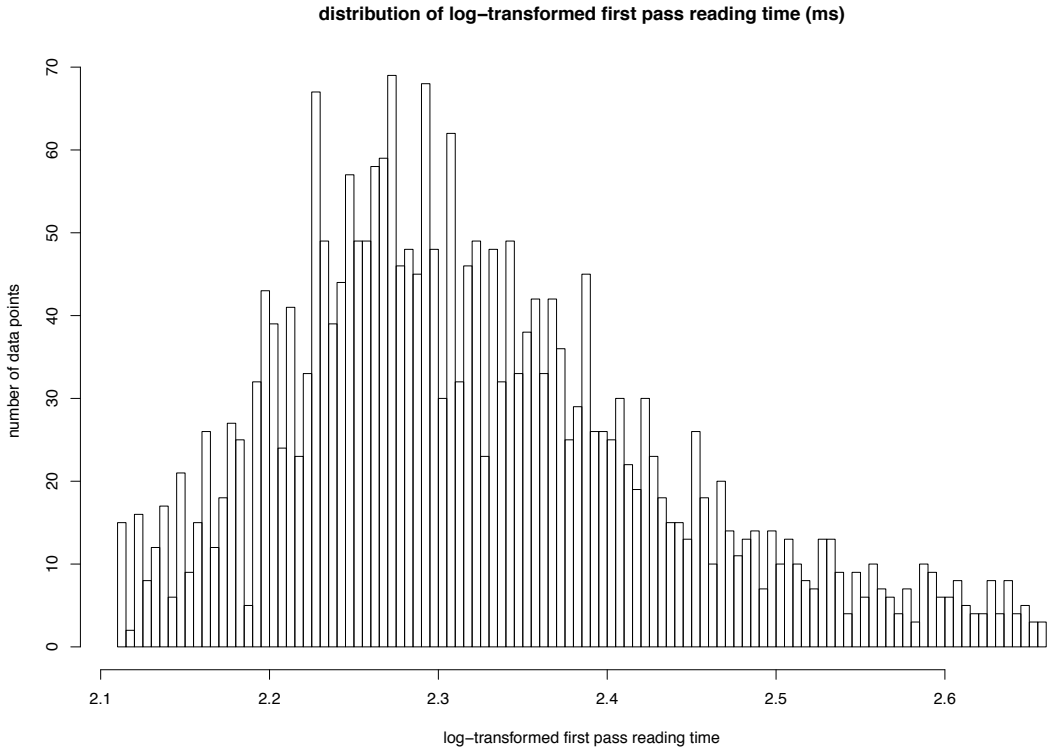
Another transformation that is often used in mixed models is *scaling*. Often scaling is performed with a z-transformation. The values of variable are centred around the mean and divided by the standard deviation:

$$x_{scaled} = \frac{x - mean(X)}{\sigma} \tag{2.7}$$

[7] Self-paced reading is a method to record reading time. A participant is reading a text from a computer screen. The text is presented bit by bit, most often in bits of one word but the researcher is free to choose their own units. It is the participant themself that is in charge of displaying the next bit of text by clicking on a button when they are done reading. So, the self-paced reading time then corresponds to the time that it takes the participant to click (Rayner, 1998).

FIGURE 2.4: The log-transformation of first pass reading times of our data from the Dundee Corpus.





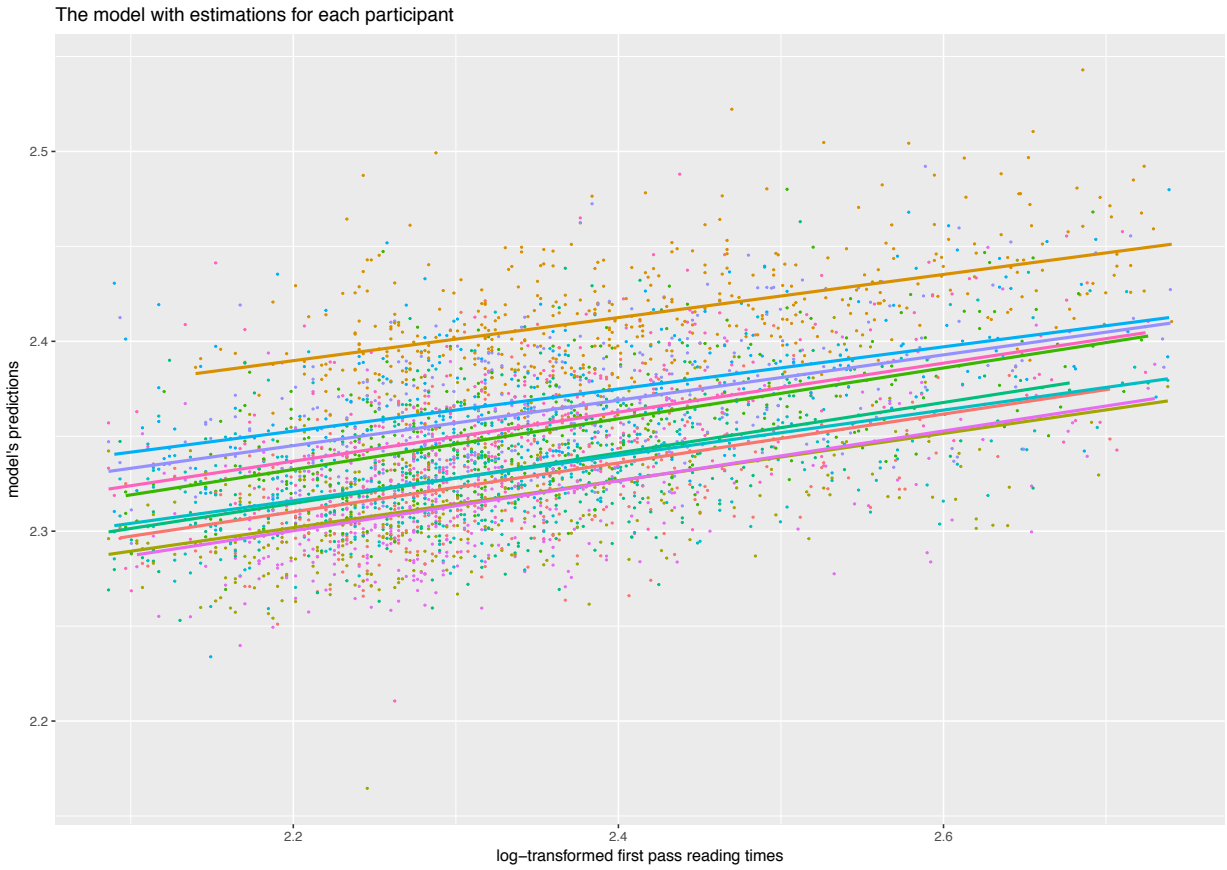
Mixed models with scaled predictor factors are on average more stable and the estimates are less dependent from each other (Field, 2009).

Fixed and Random Effects

Linear mixed effect models estimate a linear function to predict an outcome variable. 'Mixed' means that a model incorporates both fixed and random effect terms (Bates et al., 2015). Fixed effects are expected to influence the outcome variable and random effects represent groupings of data points than can influence the estimation of the fixed effects. Let us explain this by an example cited by Bates et al. (2015). In a study about sleep deprivation lasting several days, on which every day the reaction time of participants is measured, it is expected that the number of days of sleep deprivation has an influence on the reaction time of the participants: more sleep deprivation leads to longer reaction times. The number of days of sleep deprivation is therefore a fixed effect, expected to have a direct influence on the predicted variable. On the other hand, the data points (reaction times of participants) can be grouped by the participants: every participant has their own group of data points. It could be expected that each participant is different. Even though probably every participant gets higher reaction times the longer they are sleep deprived, some participants are in general quicker than others. Moreover, sleep deprivation may affect each participant in a different way. Some participants would only be bothered by it a little and others a lot. The participants in this example constitute therefore a random effect. Individual differences have an influence on how strong fixed effects, such as the number of days of sleep deprivations, are. Participants' general reaction time can be modelled by random intercepts and their sensitivity to sleep deprivation by a random slope.

In our study, all pronoun resolution biases we incorporate in the model are fixed factors, as well as the control factors that serve to filter out noise. The participants that read the Dundee Corpus and the occurrences of anaphoric pronouns (the 1109 anaphoric pronouns of our corpus) are regrouping factors, implemented as random factors in our model. Every participant has data points that belong to them and every anaphoric pronoun instance has data points of multiple participants. We decided to model random intercepts for participants and anaphora instances. This means that we estimate that not every participant reads as fast as others in general and that not all anaphora instances in the corpus are read with the same speed. But we do not implement random slopes for participants and anaphora instances. If we wanted to do so, we should assume that not every participant reacts the same way to the subject preference, the parallel function, the distance between the anaphor, etcetera. This means that for every single one of these fixed effects, a random slope per participant has to be estimated. The same would be true for anaphora instances in our corpus. But this would make the model inestimable because of the large numbers of parameters. We therefore keep it by random intercepts for subjects and anaphora instances. In Figure 2.5 it is illustrated that random factors can capture individual differences: for reading times in the Dundee corpus, the influence of random factors is shown for the ten participants.

FIGURE 2.5: A model of first pass reading times for pronoun resolution from data from the Dundee Corpus. The ten participants in the data are visible: they all have their own color.



2.3.4 Control Factors

The role of control factors is to reduce the noise caused by low-level features that are known to have an influence on reading times. For example, the length in characters of the words matters, as well as their frequency: longer words take longer to read and frequent words are read faster. To account for the influence these factors have, they are included in the statistical model. Indeed, they are often causing so much variance, that not including them into the model could make it difficult to see the influence of the pronoun resolution biases.

2.3.5 Removal of Data Points

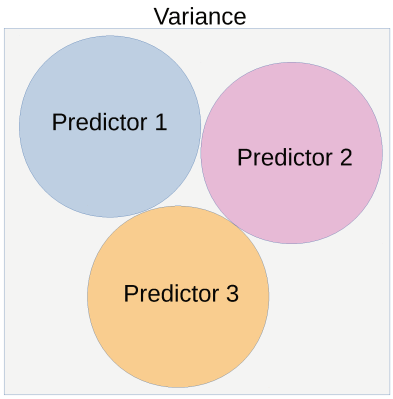
Another way to reduce noise is to remove data points from the data set. It could be decided that, for example, a comma in the zone of interest is disturbing and that repressing its influence with a control factor is not efficient enough. A solution is then to filter out data points with this unwanted property. Other researchers that worked with the Dundee Corpus (Demberg and Keller, 2008; Frank and Bod, 2011), filtered out a lot of data points with noise-introducing features.

Data points may also be removed because the reading time is deviant. For example, a reading time of 10,000 ms is abnormal. It indicates that the participant did something else than text reading. The removal of outliers can improve the fit of the statistical models and help to obtain a more normal distribution of the data. However, the removal of outliers should be done with a lot of care: long reading times can be caused by linguistic features of the text and therefore, high reading times should not be automatically removed.

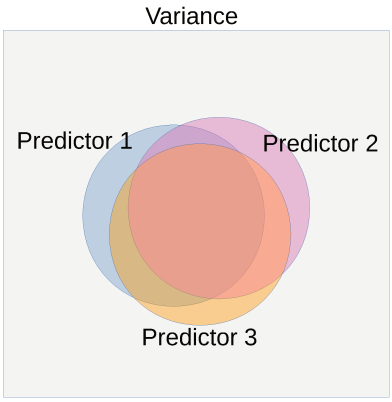
2.3.6 Multicollinearity

Multicollinearity can be a problem in the interpretation of the influence of individual predictor factors. If factors are highly correlated, they do not explain different parts of the variance in the data, but the same, as illustrated in Figure 2.6.

FIGURE 2.6: An illustration of multicollinearity. In a hypothetical space, the variance of the data is represented by a gray surface. Predictors can explain a part of the variance in the data. In the figure, this is represented by predictors covering a part of the grey surface. In the case of multicollinearity, in total, less variance is explained by the predictors because they explain the same variance in the data.



No multicollinearity: each predictor explains a part of the variance



Multicollinearity: the predictors explain the same part of the variance

If multicollinearity is high, the coefficients of correlated variables are not estimated correctly (Field, 2009). Therefore, their interpretation becomes erred. Multicollinearity can be detected by looking at the correlation matrix of the coefficients, that checks for each pair of predictors (X_1, X_2), how correlated they are. However, this method does not assess the correlations between combinations of predictors, for example the correlation between (X_1, X_2) might be small, but the correlation between ($X_1 + X_3$, X_2) might be large. A solution to detect multicollinearity in these cases, is to calculate the variance inflation factors of the predictors (VIFs). Variance inflation factors estimate how much the variance of a factor, estimated by the model, is inflated because of collinearity. When VIF is 1 for a factor, there is no correlation between the factor and the other predictors for the model. As a rule of the thumb, it

is said that VIFs greater than 4 need to be handled with care and those greater than 10 need a correction. For example, one can remove one of the correlated factors from the model.

2.4 Corpus

For our study of pronoun resolution in corpus data, we used a resource of natural text reading: the English part of the Dundee Corpus (Kennedy, Hill, and Pynte, 2003; Kennedy and Pynte, 2005), the largest eye-tracking corpus for English. The corpus counts 51 502 tokens, 9 776 types and 2 368 sentences (Barrett, Agić, and Søgaard, 2015). The corpus contains 20 newspaper articles from *The Independent*.[8] The corpus was read by ten native speakers of English with no reading or vision troubles, while their eye movements were recorded.

To this resource, we added an annotation layer in which we annotated all personal pronouns (Seminck and Amsili, 2018). This annotation layer's name is APADEC: *Anaphorical Pronouns and their Antecedents in the Dundee Eye-Tracking Corpus* (Seminck and Amsili, 2018). In total, there are 2 123 personal pronouns in the corpus, from which 1 109 were classified in APADEC as anaphoric and 1 014 as being from another category. APADEC is freely available from `http://www.llf.cnrs.fr/apadec`. In the following paragraphs, we describe the annotation process and the evaluation of the annotation layer.

2.4.1 Annotation

We chose to have the corpus manually annotated by two annotators. We searched for instances of personal pronouns using a part-of-speech tag annotation of the corpus provided by Frank et al. (2009) and transformed to Universal Dependency Part-of-Speech Tags (Agic et al., 2015) by Barrett, Agić, and Søgaard (2015). We then classified the pronouns according to whether they were anaphoric or not. The pronouns that were not classified as anaphora belonged to the following categories: deictic, non-referential, having a split antecedent, or cataphoric.

First and second person pronouns were considered as non-anaphoric, since they have a deictic function: pointing to the extra-linguistic context, instead of the linguistic context. Non-referential pronouns are pronouns that do not refer to anything, such as *it* in the sentence (6).

(6) It is raining.

Having a split antecedent means that the antecedent of a pronoun is not in just one place. An example is sentence (7). Actually, having a split antecedent is not a case of non-anaphoricity, but for practical purposes we decided not to include these pronouns in our set of anaphoric pronouns. It is for example difficult to measure the lexical frequency of the split antecedent.

(7) Florian$_{referent}$ is going on a holiday. Ernest$_{referent}$ will come too. They$_{resolve}$ go by car together.

[8] It seems rather to be 20 sessions of reading, each session containing multiple articles, but no information about this is conserved in the description of the corpus.

We excluded cataphora because we assume that the resolution process is different from anaphora. Cataphoric pronouns are pronouns that appear before their referent, as illustrated in example (8).

(8) When he$_{resolve}$ was young, Ernest$_{referent}$ lived in the south of France.

For anaphoric pronouns we annotated the closest mention of the antecedent, by taking its span of words. This is illustrated in Table 2.3.

TABLE 2.3: An antecedent annotation example from the corpus

word nb	word form	POS	antecedent
1800	if	[IN]	
1801	the	[DT]	
1802	voters	[NNS]	
1803	did	[VBD]	
1804	not	[RB]	
1805	care	[VB]	
1806	about	[IN]	
1807	that,	[DT, ,]	
1808	they	[PRP]	1801-1802
...	...	...	...

The annotation was done in the following manner:

- Annotator 1 annotated the entire corpus.
- Annotator 2 was instructed by Annotator 1 and annotated separately 36 232 words of the corpus.
- Annotator 1 and 2 compared their annotations and decided upon all cases they did not agree on.
- Annotator 1 corrected the $\sim$ 14 000 remaining words of the corpus for mistakes.

2.4.2 Evaluation

To evaluate the inter-annotator agreement for distinguishing anaphoric from non-anaphoric pronouns, we used Cohen's κ, a measure of agreement adjusted for chance (Cohen, 1960; Artstein and Poesio, 2008). Cohen's κ was 0.88 (Pedregosa et al., 2011) for the 36 232 words Annotator 1 and 2 annotated separately. This indicates a very good agreement. Second, we evaluated the identification of antecedents. Because this task consisted in giving the span of words that corresponds to the antecedent, we can say that for this task, there is no obvious set of labels available, and therefore a measure as Cohen's κ is inadequate. Even, if it were possible to consider every possible span of words in a text as a potential label, this does not resolve the problem that the spans two annotators identify can overlap, without being exactly the same. A metric that can handle non-categorical data is Krippendorff's α (Krippendorff, 1980). The value of α is 1 minus the ratio between observed and expected agreement. Passonneau (2004), Passonneau (2006) and Artstein and Poesio (2008) propose various ways to adapt Krippendorff's α to the situation in which labels are sets. Disagreement can be quantified by various distance metrics to account for

set similarity (see Table 2.4). We applied the α-metric to our data using the implementation provided in the NLTK-library (Bird, Klein, and Loper, 2009), considering antecedent spans as sets of words. The scores are given in Table 2.4.

		α
binary distance	$d_b =$ 0 if $s_1 = s_2$ 1 if $s_1 \neq s_2$	0.71
Jaccard distance	$d_j =$ $1 - \frac{\lvert s_1 \cap s_2 \rvert}{\lvert s_1 \cup s_2 \rvert}$	0.78
MASI-distance	$d_M = d_j \cdot M$, with $M =$ 0 if $s_1 = s_2$; $\frac{1}{3}$ if $s_1 \subset s_2$ or $s_2 \subset s_1$; $\frac{2}{3}$ if $s_1 \cap s_2 \neq \emptyset$ but $s_1 \not\subset s_2$ & $s_2 \not\subset s_1$; 1 if $s_1 \cap s_2 = \emptyset$	0.75

TABLE 2.4: Values of Krippendorff's α to measure the inter-annotator agreement on the span of words for the antecedent given various distance metrics.

It is often assumed that α's > 0.67 is enough to support cautious conclusions, and in that light our annotation seems rather reliable. However, Passonneau (2006) and Artstein and Poesio (2008) warn that this is not a hard value and that it is heavily dependent on the data. We therefore also measured the reliability of the annotations by comparing both annotations to the final annotation of our corpus. In the field of anaphora resolution, the information retrieval metrics of precision and recall — see Equations (12) and (13) — are often used to measure the quality of coreference chains (Vilain et al., 1995; Artstein and Poesio, 2008). We defined the notions of *true positive (tp)*, *false positive (fp)* and *false negative (fn)*, so that they would fit our situation in which only a part of the annotated span can match the span of the final annotation (see formulas (9), (10), (11) for an explanation).[9] For every anaphor in our corpus, we thus get a score between 0 and 1 for *tp*, *fn* and *fp* and then calculate precision and recall in the usual manner. In Table 2.5, the scores of both annotators can be found. Both annotators seem to obtain good scores. The differences between the scores of the two annotators can be explained by the fact that Annotator 2 sometimes annotated the right antecedent, but not the closest and that she is an undergrad student in humanities, whereas Annotator 1 is the author of this thesis.

(9) $$tp = \frac{|\text{ann} \cap \text{gold}|}{|\text{gold}|}$$

(10) $$fn = 1 - \frac{|\text{ann} \cap \text{gold}|}{|\text{gold}|}$$

(11) $$fp = \frac{|\text{ann} - \text{ann} \cap \text{gold}|}{|\text{ann}|}$$

(12) $$\text{Precision} = \frac{tp}{tp+fp}$$

(13) $$\text{Recall} = \frac{tp}{tp+fn}$$

[9] In the formulas we refer to our final annotation with the word *gold*.

TABLE 2.5: Evaluation of the identification of antecedents using precision and recall.

	precision	recall	F_1
Annotator 1	0.95	0.92	0.93
Annotator 2	0.81	0.81	0.81

2.5 Model

In this section, we present experiments for examining the reading times of anaphoric pronouns and the words around them of the Dundee Corpus. For every pronoun, a window of six words is defined, with the pronoun in the second position (see (14-a) and (14-b) for example).

(14) a. when **they** are at great risk
b. but **it** would seriously degrade the

For every position in this window, a linear mixed effects model is estimated. The reason for choosing a large window is that there is a possibility of parafoveal reading and spill-over effects. It may seem a very large zone, but one has to keep in mind that it remains unclear where to expect the influence of pronoun resolution and that other researchers, such as Von der Malsburg (2018) also choose this window.

The experiments here serve two goals. A first goal of our research is to seek confirmation for a maximum of factors of influence described in a large psycholinguistic literature (see Section 2.2). We studied the distance between the pronoun and its antecedent (Clark and Sengul, 1979; Ehrlich and Rayner, 1983), the frequency of the antecedent (Shillcock, 1982; Pynte and Colonna, 2000; Van Gompel and Majid, 2004), the subject bias (Broadbent, 1970; Clancy, 1980; Frederiksen, 1981), grammatical parallelism (Maratsos, 1973; Sheldon, 1974; Smyth, 1994) and the first mention bias (Gernsbacher and Hargreaves, 1988).

A second goal of our research is to see whether our corpus can be used as a means to discover new factors and to study less well-known factors. We studied for example the difference between different genders and number of anaphoric pronouns. We wanted to study whether there was a difference in resolution time between masculine, feminine, neutral and plural pronouns. Finally, we also tested if the length of the antecedent phrase in words had an influence on the reading time of the pronoun, because according to Accessibility Theory (Ariel, 1991), more salient discourse referents are marked in a more compact way.

2.5.1 Implementation Details

For this first exploration of the corpus we choose to study *first pass reading time*, the sum of the durations of the fixations on the region before leaving the region (see Section 2.3.2 for a more detailed explanation and an example). Before we did the modeling, we followed previous studies on the Dundee Corpus (Demberg and Keller, 2008; Frank and Bod, 2011) to clean up the data. Words that were presented first or last on the line, that were not fixated, that had punctuation attached, that had clitics attached, or that contained more than one capital letter were excluded from

the data. We also cut off outliers, by cutting the data at minus one standard deviation at the left of the distribution and plus three standard deviations at the right of the distribution. We made this choice, because according to the qq-plot diagnostic, this gave the best model.

For the different regions we had between 2 593 data points (pronoun) and 4 173 (one word after the pronoun) data points left. As preprocessing we applied a log-transformation of base 10 on the reading times and scaling by a z-transformation on numeric variables. All data points including the value None were ignored.[10]

Reading Zones

We defined six zones of interest starting from one word before the pronoun up to four words next to the pronoun. Each zone corresponds to a mixed-effects model from the `lme4` R package (Bates et al., 2015). The model contains three types of factors: random factors, control factors and pronoun resolution factors. Control factors are always local to the reading zone. So, when we take the string given in (15) as an example of a window containing the six zones of interest, all control factors are calculated for each zone separately. On the other hand, pronoun resolution factors can be understood as non-local. These factors always concern the relation between the pronoun and its antecedent. Therefore, even if the reading time of *Irish* is modelled, a factor like *distance between the pronoun and its antecedent* still applies to the pronoun **she** and its antecedent.

(15) And, **she** added, the Irish have

Random Factors and Control Factors

We modelled each of the 1 109 pronouns and each of the 10 participants with random intercepts. Following other studies (Demberg and Keller, 2008; Frank and Bod, 2011) on the Dundee Corpus, we used the following control factors:

- fProba: forward probability from an n-gram model, obtained from the data distributed by Frank and Bod (2011).
- bProba: backward probability from an n-gram model, obtained from the data distributed by Frank and Bod (2011).
- length: the length in characters of the zone of interest.
- log-freq-dundee: log-frequency of the word in the zone of interest in the Dundee Corpus.
- log-freq-bnc: log-frequency of the word in the zone of interest in the British National Corpus.[11]
- prev-log-freq-dundee: log-frequency of the word previous to the zone of interest in the Dundee Corpus.

[10] A data point can include the value of None when a parameter cannot be estimated for it. For example, when there are no fixations on a word, its first pass reading time has the value of None.

[11] The BNC counts were taken from:
`https://www.kilgarriff.co.uk/bnc-readme.html`.

- prev-log-freq-bnc: log-frequency of the word previous to the zone of interest in the British National Corpus.
- launch-pos-first-fix: the launch position of the fixation. Gives the distance in letters from the previous fixation (negative for reading from left to right and positive for reading from right to left).
- land-pos-first-fix: the landing position of the fixation. The n^{th} letter of the zone on which the first fixation landed.
- sup: syntactic surprisal, calculated by a probabilistic context free grammar parser and distributed in the data of Frank and Bod (2011).

Pronoun Resolution Factors

In opposition to the control factors, the pronoun resolution factors are constant over the six regions, because they consider the anaphoric relation that is not marked specifically on one of the words of the six regions. The pronoun resolution factors correspond to the biases we discussed in section 2.2. We introduce them and their implementations in the following list. We also discuss the new factors we studied: the gender and number of the pronoun and the length of the antecedent.

- dist-ant-begin: the distance in words between the antecedent and the beginning of the text. This factor is inspired by the first mention preference, described in Section 2.2.3.
- dist-ant-ana-words: distance in words between the anaphor and its antecedent. This factor is inspired by the short distance preference described in Section 2.2.4.
- log-freq-dundee-head-ant: the log frequency of the syntactic head of the antecedent in the Dundee Corpus. This factor measures the frequency specific to the Dundee Corpus. See Section 2.2.5 for details about the influence of frequency.
- log-freq-bnc-head-ant: the log frequency of the syntactic head of the antecedent in the British National Corpus.[12]. This factor measures the frequency of the word in the English language in general. See Section 2.2.5 for details about the influence of frequency.
- syntactic-role-ana: the grammatical function of the pronoun. To calculate this, we took the universal dependency annotation of Barrett, Agić, and Søgaard (2015) and regrouped all the different functions into three main functions: subject, direct object or other. This was necessary to overcome the problem of data-sparsity, which is very large if the categories that are initially in the data are maintained. This variable is encoded by dummy encoding.
- syntactic-role-head-of-antecedent: the grammatical function of the head of the antecedent (subject, direct object or other). Just as for the grammatical function of the pronoun, we regrouped the functions of the head of the antecedents by three syntactic categories, again because of the problem of data sparsity. This factor can tell us more about the subject preference (see Section 2.2.1). This variable is encoded by dummy encoding.

[12]See footnote 11.

- parallel-func: the parallel function. After regrouping all the different categories into the three main functions (subject, direct object or other) to overcome data-sparsity, the syntactic function of the antecedent and that of the pronoun are compared. If they are the same, the value of the factor is True, otherwise False. A description about the parallel function preference can be found in Section 2.2.2.

- form pronoun: this is a categorical variable with the four values: *he, she, it, they*. When the form of pronoun is *he*, or *him*, the value is *he*, when it is *she*, or *her*, it is *she*, when it is *they*, or *them*, the value is *they* and of course, the value is *it* when the form of the pronoun is *it*. This variable is encoded by dummy encoding.

- length-ant: this factor is suggested by *Accessibility Theory* that states that referents that can easily be retrieved from memory, because they are salient, are expressed in a more compact way than referents that are difficult to retrieve from memory (Ariel, 1991). The length of the antecedent (number of characters) measures the compactness of the expression of the antecedent.

2.5.2 Results

In this section, we present the outcome of the models and how it relates to the psycholinguistic literature presented in Section 2.2. The effects are rather small and we therefore will discuss in more details the meaning of the results in Section 2.5.3.

The results of the models are reported in Table 2.6. Each column under a region — 0 for the pronoun — represents a model for that region. The numbers for each factor are the model's coefficient estimate for that factor. A positive estimate indicates that a higher score for the factor increases the reading time and a negative score indicates a shorter reading time. Thus, positive coefficients indicate higher cognitive load and negative coefficients lower cognitive load.

Some interesting patterns can be read from the results about the pronoun resolution factors: we will shortly discuss them below.

Distance Factors

Let us first have a look at the distance factors. It seems that a higher distance between the antecedent and the beginning of the text matters: antecedents early in the text are retrieved faster than those further in the text (see Table 2.6, region 1). This could be explained by mechanisms such as the *first mention preference* that attributes greater saliency to early mentioned antecedents.

There is also an effect of the distance between the pronoun and its antecedent. That longer distances gives less reading time is suggested by the distance factor represented by the number of words between the pronoun and the antecedent (region -1). This is not in line with the psycholinguistic literature presented in Section 2.2.4. However, an effect in the opposite direction is suggested for the same factor in region 3. So, this second result supports the conclusions of psycholinguistic research, cited in Section 2.2.4. We can suggest two explanations for this: either we can interpret this pattern as a delay in processing when the distance is long, not in the total cognitive load. It would suggest that for long distances, first there is a speed-up and then the pronoun is processed later, whereas for shorter distances the pronoun

TABLE 2.6: The modelling results of the six regions of interest. For every zone, from one word before the pronoun upto four words after the pronoun, the coefficients of the linear mixed models are reported, as well as the level of significance. Please see Section 2.5.3 for a more detailed discussion about these levels.

Region	-1		0		1		2		3		4	
(Intercept)	2.4E+00	***	2.2E+00	***	2.3E+00	***	2.3E+00	***	2.3E+00	***	2.3E+00	***
Control Factors												
fProba	-1.4E-03		7.9E-03	.	-4.8E-03		6.4E-04		-1.0E-03		-2.3E-03	
bProba	-2.1E-03		3.4E-03		1.9E-02	***	5.0E-03		5.0E-03		-3.7E-03	
length	-7.4E-03		2.1E-03		-2.1E-03		1.3E-03		1.2E-02	*	-5.3E-03	
log-freq-dundee	-1.1E-02	*	-6.6E-02	.	-9.0E-04		4.1E-03		1.9E-03		-9.0E-03	.
prev-log-freq-dundee	-2.2E-03		1.1E-03		-6.6E-03		1.1E-02	.	-1.2E-03		1.1E-03	
log-freq-bnc	3.2E-03		1.6E-01		-9.0E-04		-8.7E-03	*	5.8E-03		1.2E-03	
prev-log-freq-bnc	-5.5E-02	**	-1.8E-03		-6.6E-03	.	-1.2E-02	**	-2.2E-03		-1.2E-03	
launch-pos-first-fix	-6.0E-03	*	-4.6E-03	*	-6.7E-03	***	-4.9E-03	*	-3.7E-03	.	-3.2E-03	
land-pos-first-fix	2.1E-03		1.7E-03		-1.0E-03		2.1E-03		-8.3E-04		-4.0E-03	
sup	3.8E-03		8.6E-04		1.3E-03		1.1E-03		8.3E-03	*	-1.2E-03	
Distance Factors												
dist-ana-ant-words	-4.1E-03	*	2.2E-03		6.4E-04		1.1E-03		4.2E-03	*	2.4E-03	
dist-ana-begin	-3.3E-03		-2.4E-04		4.7E-03	*	7.6E-04		-9.2E-05		-6.3E-04	
Frequency Factors												
log-freq-dundee-head-ant	1.8E-03		2.3E-03		2.8E-03		2.7E-03		-5.5E-03		-8.5E-03	*
log-freq-bnc-head-ant	-2.5E-04		1.4E-03		-4.8E-04		-7.2E-05		3.8E-03		5.6E-03	
Grammatical Function Factors												
syntactic-role-anansubj	6.4E-03		8.1E-03		-7.1E-03		-2.0E-02	*	-7.5E-03		-2.7E-03	
syntactic-role-anaother	9.3E-04		5.3E-03		4.2E-04		-3.3E-03		-1.2E-02		-9.5E-03	
syntactic-role-head-of-antecedentnsubj	7.2E-04		3.7E-03		1.4E-02	.	4.8E-03		1.1E-02		1.1E-02	
syntactic-role-head-of-antecedentother	1.6E-03		-1.3E-03		2.0E-02	*	6.7E-04		1.3E-02		-6.6E-03	
parallel-funcTRUE	-5.6E-03		-1.3E-02	.	-2.6E-03		-1.2E-04		-2.8E-03		-1.1E-02	
Gender Number Factors												
form pronoun they	1.4E-03		1.5E-02		-1.9E-02	*	-8.5E-03		-1.2E-02		7.4E-03	
form pronoun it	-1.0E-02		2.5E-02	*	-3.3E-03		-4.9E-03		1.7E-03		1.2E-02	
form pronoun she	-1.2E-02		-1.7E-02		-7.6E-03		7.8E-03		-1.2E-02		-1.7E-03	
Accessibility Theory Factors												
length-ant	1.0E-03		7.2E-04		3.6E-04		1.4E-03	.	-1.1E-03		-2.1E-04	

Significance: *** for $p < 0.001$; ** for $p < 0.01$; * for $p < 0.05$ and . for $p < 0.1$.

is processed immediately. A second, and maybe more plausible explanation, is that one of the results, likely the one in region -1, is the result of a statistical error. Indeed, in Section 2.5.3 we will discuss the role of statistical errors in the presented design.

Frequency Factors

When we have a look at our two frequency factors, we see that only the factor that takes into account the frequency of the head of the antecedent in the Dundee Corpus is significant (region 4) while the frequency of the head of the antecedent in the British National Corpus does not show any significant effects. This result has a negative estimate, thus indicating that a higher frequency in the Dundee Corpus leads to faster reading times. The part of the result that suggests that higher corpus frequency of the antecedent gives shorter reading times is partly consistent with the psycholinguistic literature (see Section 2.2.5), but it would be expected that the result should also show on the parameter that measures frequency in the British National Corpus. It could be the case that the effect that shows that the frequency of the antecedent is relevant in the Dundee Corpus, actually measures how often the entity has been mentioned so far and that that is the reason that only the frequency in the Dundee Corpus matters to the model.

Grammatical Function Factors

When we consider grammatical function, the direct object was encoded as the baseline of the dummy encoding. We see that when the pronoun has a subject function, there is a speed-up in region 2, in line with the saliency account, suggesting that pronouns are expected in the subject position. Finally, when looking at the grammatical function of the antecedent, we see that for antecedents that have a function other than subject or object, processing is slower (region 1). This last result is also in line with the saliency account that states that direct objects are easier to retrieve from memory than other grammatical functions, except for the subject function.

Gender and Number of Pronoun

When we look at the different pronouns for different numbers and genders (*he*, *she*, *it*, *they*), we see some interesting results. Here, the baseline of the dummy encoding is the masculine singular (*he* and *him*), so these results should be interpreted against this category. We see that the neutral gender (*it*) leads to more reading time (region 0). This effect can be explained by the fact that *it* can be anaphoric and pleonastic, but all the occurrences we studied were anaphoric. Hence, the increased reading time can come from the decision readers have to make whether the pronoun is anaphoric or not. On the contrary, we found that the plural pronoun (*they*) leads to a speed-up (region 1) compared to the other pronouns. We think that *they* might be less ambiguous than other pronouns, having perhaps frewer compatible antecedents.

Length of the Antecedent

For the factor of the length of the antecedent, inspired by *Accessibility Theory* (Ariel, 1991), we did not find any effect. But we should note that the Accessibility Theory makes more predictions about production than about comprehension. It states that

inaccessible referents are realized with a more detailed, and hence longer, nominal phrase. In our experiment, we tested whether it was also true that an antecedent that was longer was easier or more difficult to retrieve from memory at the moment the pronoun was read. We thus only tested whether the length of the referential form of the antecedent had an influence on the reading of pronouns and did not give an exact implementation of the theory.

2.5.3 Discussion

The model presented in this chapter shows some interesting results and it seems that the study of pronoun resolution on corpus data can lead to new insights. We also observe that most factors only show an effect in one zone of interest and that the effect of significance appear only at the 0.05 level.

In this section, we want to take stock of the experiment and discuss what we can learn from it and how it could be improved in future work. We first discuss where the effect of pronoun resolution shows in natural data. Then, we discuss the amount of multicollinearity in the models. We finish the section by discussing various parts of the method that need improvement: the many zones tested, the random variable structure, the encoding of categorical variables and the number of predictor variables.

Where does the effect of pronoun resolution show?

The results from this experiment can give some insights on where the effect of pronoun resolution can show in natural reading data. We will discuss all the zones of interest $[-1, 4]$.

Only the effect of distance shows in the region before the pronoun, but the effect does not go into the expected direction. Psycholinguistic literature predicts that more distance should lead to more reading time, but the effect in this region suggests the opposite. We therefore keep in mind that it might be a statistical error. The absence of expected effects in this region could suggest that the people do not yet experience the cognitive load of pronoun resolution when they read the word before the pronoun.

When we look at the region of the pronoun itself (zone of interest 0), we see a delay when the pronoun is neutral (*it*). As we suggested in the previous section, this can be because the word '*it*' is ambiguous between an anaphoric pronoun and a non-referential pronoun. Hence, the result could reflect this effect instead of the effect of pronoun resolution. In general, it is not surprising that the effects do not show often in the zone of the pronoun (region 0), because only about 40% of the pronouns of the Dundee Corpus are fixated.

When we look at region 1, the word after the pronoun, we see that most of the pronoun resolution related effects showed here. Three effects are in this region and the two effects described in the literature (the distance effect and the syntactic role of the antecedent) go in the expected direction. We think therefore that this region is the most interesting region to study when we examine first pass reading time.

As for the second word up to the fourth word after the pronoun, there is only one effect per region that shows. The effects in region two and three go into the direction that is expected according to the psycholinguistic literature, and the effect of the

frequency of the antecedent found in region four is compatible with the psycholinguistic literature, but not predicted, because the literature predicts an effect of the overall lexical frequency of the antecedent in the language, whereas an effect of the lexical frequency of the antecedent in the Dundee Corpus was found. The situation of only one effect per region leads us to think of two hypotheses. The first hypothesis is that pronoun resolution has a long spillover effect, caused by features that take a long time to retrieve from memory and this is what causes longer spillovers. The second hypothesis is that these results could be statistical errors. Indeed, the method of modelling presented might be sensitive to false positive results, which we will explain in the next paragraphs.

Multicollinearity of the Factors

To address the problem of multicollinearity of the factors in our mode, we calculated the variance inflation factors of the predictor variables. We used the function implemented in R by Austin Frank, [13] because this function is compatible with the lmer function, whereas other VIF-functions are not compatible with a random variable structure. On the one hand, for our six models, the VIF values were all below 4 for our pronoun resolution variables, except for the pronoun zone, where the forms of the pronoun *she* and *they* showed VIFs of 5. This also suggests that zone zero is not a good region for measuring first pass reading time for pronoun resolution. On the other hand, there were high values for some of the control variables. Word length, word frequency and the forward and backward n-gram probability were correlated. This correlation is expected, but as this study does focus on the pronoun resolution factors and does not try to interpret the control variables, we do not consider it as a major problem for the interpretation of pronoun resolution biases. It suggests nevertheless that the long list of control variables is not needed and that it can complicate the model unnecessarily.

The Problem of Multiple Zones of Interest

The problem to test multiple zones of interest is that the risk of false positive results inflates. For each test (in our study each region), a risk of 5% error is accepted. However, when multiple tests are done, for example when multiple reading measures are used, or when multiple zones of interest are defined, the risk of false positives increases (Von der Malsburg and Angele, 2017). In the case where four tests are performed on the same data, the level of α is actually not 0.05, but $1 - 0.95^4 = 0.185$, under the assumption that there is no correlation between the outcome of the four tests (Von der Malsburg and Angele, 2017). However, in the case of eye-tracking, there is a correlation between different reading measures and in a less extent between reading times of two consecutive zones of interest, giving — in the case of four tests — an α-value between 0.05 and 0.185. But in the past, many researchers wrongly assumed that because of the correlations within reading time data, the false positive level was not inflated much, or at least not enough to matter. For our experiment it is thus the case that, theoretically, the α-value lies between 0.05 and $1 - 0.95^6 = 0.265$. This means that out of the eight effects we found, we could expect to get two false positives.

[13] https://github.com/aufrank/R-hacks/blob/master/mer-utils.R

A solution to this problem, recommended by Von der Malsburg and Angele (2017), is to use a Bonferroni correction (Bonferroni, 1936). This correction consist of dividing α by the number of tests carried out, in our case six: $0.05/6 = 0.0083$. It must be clear that all of the effects found in our data set are significant on the 0.05 level. Therefore, because none of them were significant on the 0.01 level, nor on the 0.001 level, none of the results would turn out significant if we used the Bonferroni correction. Von der Malsburg and Angele (2017) also underlined that the Bonferroni correction goes with a loss of statistical power: they estimated that it reduces statistical power for eye-tracking data by approximately 20%. This makes small effects — that is to say effects that only make a couple of milliseconds of change in reading times —, even if they were true, more difficult to detect.

Even though these considerations show that we should take the results of our model with extreme care, we think that it is a start to study pronoun resolution on corpus data. The good news is that most of the effects seem to concentrate in region 1 (the word after the pronoun). A very simple solution for further studies would be to study only this region and leave it at that. However, the results on some features of the antecedent suggest that the spillover effect for the frequency of the antecedent might be lost if we only looked at region 1. Indeed, in our data the effect showed very late in region 4.

An option we have explored is the fusion of multiple words into one zone of interest. We thought that in this way, we could still account for spill-over effects, but without having to reduce statistical power with the Bonferroni correction. But unfortunately, no effect whatsoever showed with this type of reading zone. It has been underlined that most effects show only in one zone and that it cannot be detected when the zones are agglomerated together, because they only appear very locally. Another reason is that zones of multiple words have less good control factors. For example, the lexical frequency of words is only measurable on a word level.

A third solution that could be considered would be to keep multiple regions, but reduce the number drastically: as there is no effect in zone -1, this zone could be cut off. Moreover, there is neither an effect in zone three. In zone four there is only one effect, so we could cut after zone two, leaving us with zone 0, 1, 2. This would give a Bonferroni correction of $0.05/3 = 0.017$, which would already be more workable. Ideally, the experiment would be repeated on a data set that allows for more statistical power. A way to obtain this is to work with a data set with more subjects. It is not necessary to have more items: the Dundee Corpus has more than 1 000 anaphorical pronouns.

The Problem of Random Intercepts Only

An appropriate random effect structure is important to prevent from false positives. We modelled random intercepts for participants and pronouns of the corpus. However, various studies show that using only random intercepts and no random slopes can inflate the risk of false positives (Matuschek et al., 2017). Therefore, it would be more careful to also introduce random slopes for the anaphora factors. However, the number of factors is so high in our model, that introducing random slopes for participants — for pronoun occurrences it might be not necessary, as the number of pronouns is very high — would for sure make the model fail to converge and to be estimated.

The Encoding of Categorical Variables

We saw that the dummy encoding used in this study has the downside of only making a limited number of comparisons between categories of a categorical variable: every category is only compared to the category in the baseline level. A solution would be to employ one-hot encoding. But we should underline that this can augment the data sparsity and that it does not do a pairwise comparison, but a comparison of every category against all other categories. If this study is redone on another corpus that would allow for more statistical power, we would therefore plead for a one-hot encoding scheme, but without automatically including all categories of a variable. So, for example, we could just transform the variable of the syntactic function into a variable that states whether something is a subject or not. We would especially choose the subject function, because the subject preference takes an important place in the psycholinguistic literature.

Number of Predictor Variables

The number of predictor variables is very high in our model. We saw that the multicollinearity of the pronoun resolution factors was not worrisome, but the multicollinearity of the control factors was on the high side. This means that in follow-up studies, it would probably be better to pick them with more care. The great number used here, inspired by previous study on the Dundee Corpus, is not necessary and the effects of low-level factors can probably be barred efficiently without having so many.

A second reason to be critical at the number of predictor variables is that a low number of predictor variables gives more power to the statistical model — that is to say that models that have fewer predictor variables have more chance to detect effects (Field, 2009). Moreover, fewer variables would make it more efficient to compute the models.

2.6 Conclusion

The model presented in this chapter is the first model designed to study human pronoun resolution in natural text. Some factors known from psycholinguistic studies also show in these data: the distance between the pronoun and its antecedent, the frequency of the antecedent and the grammatical role of both the pronoun and its antecedent. The problem remains, however, that it is not clear where the effect of pronoun resolution shows, the time course of resolution seems to be smeared over a very large region. To account for this, we needed multiple testing. To blend several words into one reading region is not a solution, because effects show at different time points and there is no effect that remains for two following words. Some factors seem to cause greater spillover effects than others, even though we need to keep in mind that these effects might actually be tested at an unacceptable high error-level and could therefore be false positives. The data of the Dundee Corpus does not have enough statistical power to enable the needed statistical correction, in order to lower the α error-level. Therefore, the results should be interpreted with the greatest care. For these reasons, we would like to underline that our study serves rather as a first exploration of reading time corpora to investigate pronoun resolution than as a hard proof of pronoun resolution effects in natural text reading.

The results obtained with our experiments show that it is likely that effects from psycholinguistic literature will show in natural reading data and that reading time corpora can be a valuable resource for this study, but that a lot of statistical power is needed to get statistically valid studies, so the error-level of the data can be corrected. To get more certainty about the effects that show in our study, it would be necessary to repeat this study on another data-source. This possibility will be discussed further in the perspectives presented at the end of this thesis (see Chapter 6).

What we learned from our study is that because of the low number of fixations on the pronoun and the absence of sensible results in the region of the word before the pronoun, these regions are probably unsuited to study reading times of pronouns. When pronoun resolution is studied on first pass reading times, the effects show as a spillover effect. Because of the inconclusive result in region 4, we suggest that further studies limit themselves to the study of the first word after the pronoun up to the third word.

Another important lesson we learned during this study is that linear modelling is not ideal to study reading time data. Even though the log-transformation is commonly used in many studies reported in the literature, parametric tests show that the log-transformation is often not good enough. This prevented us in the present framework to investigate other reading times that reflect later processing, such as the regression path duration or the total reading time. We simply could not get a model that fulfilled the parametric assumptions. We recommend therefore that for further studies, an adapted distribution is directly employed, such as the Gamma distribution. However, when researchers want to exploit other reading time metrics, they will also need a statistical correction because of repeated testing. The search for the location where the effect of pronoun resolution shows also starts all over again.

In the following chapters of the thesis, we present other models of pronoun resolution with data from eye-tracking corpora. But in these studies, we kept all the shortcomings of the present study in mind to avoid statistical problems and modelling problems. These studies are nevertheless inherently different from the present study, because they do not test all anaphora factors separately, but assume one underlying explanatory variable.

Chapter 3

Simulation of Human Behaviour

3.1 Introduction

This chapter continues to investigate pronoun resolution biases. In Chapter 2, we tried to see whether resolution biases could be attested in uncontrolled data. In this chapter, we will look at them from another point of view: we want to see whether the biases attested in human data also show in automatic systems of pronoun resolution. In other words: can automatic systems be accurate models of human behaviour on pronoun resolution? If the answer to this question is positive, it means that automatic systems can be used as tools to investigate human pronoun resolution biases. The advantage of these systems is that they are designed to be run on corpus, which makes them robust. Investigating the potential of such systems to be used to simulate human behaviour is important to move towards testing theories of pronoun resolution on corpus data. Please note that experimental results in this chapter have been published as a conference proceedings' article (Seminck and Amsili, 2017).

In this chapter, we will only focus on two biases already introduced in the previous chapter: the bias to resolve a pronoun to the syntactic subject (see Section 2.2.1 from Chapter 2) and the bias to resolve the pronoun to an antecedent that occupies the same syntactic position (see Section 2.2.2 from Chapter 2). The reason why we choose these two biases is that they can make conflicting predictions when they are applied as resolution heuristics. This is especially the case for pronouns that are not in the subject position, such as in (1).

(1) Florian sees Ernest when Anna meets him$_{resolve}$ outside.

As said in Chapter 2, the psycholinguistic literature reports different results for these two biases: Crawley, Stevenson, and Kleinman (1990) only found an influence of the subject bias and Smyth (1994) found an influence of both biases. It is suggested that the conflicting results emerge from differences in syntactic structure in the experimental items used for the studies (Smyth, 1994). We think that an automatic pronoun resolver can provide an interesting framework that investigates how these biases work and explain the different results from the psycholinguistic studies. In our automatic model, it is necessary to give a precise and exact implementation of the biases, because the model needs to determine automatically the cases where they apply. For the biases we decided to study, it is crucial to specify the type of syntactic analysis that is applied. Indeed, where exactly a parallel syntactic function is detected depends greatly on how syntactic structure is analysed. Some frameworks provide a more fine-grained analysis then others which can lead to the situation in which according to one framework structures are parallel, whether in another

framework that employs a more fine-grained analysis, the structures are analysed as being different. Therefore, a computational implementation leads to a higher degree of precision of the definition of biases than in most psycholinguistic experiments. We show that a very precise level of the formulation of the pronoun resolution biases can provide new insights into how they work. We also show that it can give rise new research questions.

For our experiment, we implemented a very simple system for pronoun resolution based on some basic features used in almost all modern systems of coreference and pronoun resolution in the NLP-community. Our system works with a logistic regression function and is trained on a corpus annotated with coreference chains. We chose this technique because the coefficients of the system can be interpreted easily and provide insights into what the system learned from the corpus data.

To compare the biases that the automatic pronoun resolution system learned to the biases attested in the psycholinguistic experiments, we ran our model on items from those experiments. We used the items of experiments in which participants had to choose antecedents for ambiguous pronouns. We compared the human choices to the model's choices and concluded that they were quite similar. This shows that the model learned the biases accurately from corpus data. Hence, automatic systems can serve as a model to study human pronoun resolution.

This chapter also provides an overview about the work on pronoun and coreference resolution in the field of NLP. Because this thesis makes use of NLP-techniques (also in Chapter 4), it is important to know what today's state of the art is and which theories and approaches have been proposed. This chapter starts with a description of pronoun and coreference resolution systems in the field of NLP in (Section 3.2). After that, we present our experiment in which we simulated the subject bias and the parallel function bias with a pronoun resolution system. The resolution system we implemented and trained on corpus data is described in Section 3.3. Eventually, we present the results of the human behaviour simulation experiment on the choice of antecedents of ambiguous pronouns in Section 3.4 and finish with a brief conclusion in Section 3.5.

3.2 Coreference and Pronoun Resolution in NLP

The topic of anaphora resolution dates back to the beginning of the field of NLP. Indeed, anaphora resolution is important for machine translation and because the field started there, anaphora resolution soon became a topic as well. In the literature about anaphora resolution, different time periods are marked by some key works. Here, we will roughly divide this 'history' into two periods of research: a period of early rule-based works that goes from the work of Hobbs (1978) until the 1990's to the work of Lappin and Leass (1994) and Centering Theory based algorithms (Brennan, Friedman, and Pollard, 1987) and a second period starting around the millennium with the start of shared tasks like MUC (*Message Understanding Conference*) and the famous work of Soon, Ng, and Lim (2001). We try to sketch the landscape of anaphora and coreference resolution in NLP, so the reader knows what type of approaches exist and how they relate to each other.

Anaphora resolution, and thus pronoun resolution, are part of coreference resolution. However, they are not exactly the same phenomenon. Coreference means that two, or more, linguistic expressions refer to the same entity. Van Deemter and

Kibble (2000) used the following equation to illustrate a coreference relation between referential expression α_1 and α_2:

$$Referent(\alpha_1) = Referent(\alpha_2) \tag{3.1}$$

For example in a sentence like (2-a), the two mentions of *Ikea* are coreferent: they refer to the same entity. But we cannot say that there is an anaphoric relationship between the two mentions, because for an anaphoric relationship it is necessary that the second mention depends on the first for its interpretation (Van Deemter and Kibble, 2000). To access the sense of the second mention, it is not needed to recall the sense of the first mention, that is to say: a resolution process is not needed. On the contrary, in example (2-b), the pronoun *it* needs to be resolved: the mention of *Ikea* must be retrieved. Recall that it is still the case that *Ikea* and *it* refer to the same entity, therefore they are also coreferent.

(2) a. Florian hates to go to $\text{Ikea}_{coreference}$, but $\text{Ikea}_{coreference}$ sells convenient storage systems.
b. Florian hates to go to $\text{Ikea}_{antecedent}$, but $\text{it}_{anaphor}$ sells convenient storage systems.

The purpose of underlining this difference between anaphora and coreference resolution is that, for a long time now, in the NLP-community little attention has been paid to it. Whereas earlier works focused more specifically on pronoun resolution, from the end of the 1990's, research in the NLP-community became increasingly focused on *shared tasks*. Shared tasks are competitions where scientists are challenged to develop the best performing system for a task, for example to group all linguistic expressions that refer to the same entity in the case of coreference resolution. Evaluation is done automatically using a scoring function. The shared task organisers provide data that the researchers use to develop their systems. Once all the systems have been sent in, a held-out data set — the test data — is used to evaluate the systems. The reason that little attention is paid to different types of reference is that it was not required by the shared tasks. Even though some critiques have been formulated by for example Van Deemter and Kibble (2000), who argued that the concept of *coreference* as is it used by the NLP-community includes phenomena that rely on distinct semantic mechanism, most systems — including today's state of the art — do not make explicit distinctions between different forms of coreference and anaphora.

In this section, we discuss the systems from the time period before the introduction of the shared tasks, the data sets used for the shared tasks as well as the evaluation metrics and today's state-of-the-art systems. The goal is to give the reader a global picture of the models built in the NLP-community, be they specifically about pronouns or coreference in general. A full description of all previous systems is out of the scope of this thesis. We refer the reader to the textbook *Anaphora Resolution* by Poesio, Stuckardt, and Versley (2016) that gives a far more extensive overview.

3.2.1 Early Rule-Based Systems

The earliest anaphora resolution systems date from the 1960's (Stuckardt, 2016). However, they were based on hand-crafted rules and only applicable in restricted domains. We refer the reader to other works (Poesio, Stuckardt, and Versley, 2016; Mitkov, 2002) for a detailed description. In this section, we rather focus on robust

systems that have been developed before the area of shared tasks that have nevertheless heavily influenced modern systems.

Hobbs' algorithm

According to Stuckardt (2016), Hobbs (1978)'s naive algorithm of pronoun resolution can be seen as one of the first works that proposes a broad coverage robust algorithm for pronoun resolution. The algorithm defines rules to find the antecedent of a pronoun using the syntactic trees of the sentence where the pronoun occurs and the preceding sentences. The algorithm searches for noun phrases (NPs) following a search path through the syntactic tree. When an NP is met, it is checked to be compatible in gender with the pronoun. It is also checked for some semantic constraints. For example, for the pronoun *it* in example (3), the NP *536* is rejected because the verb *moved* is incompatible with a date. We refer to the original article of Hobbs (1978) for the full description of the algorithm.

(3) [The castle in [Camelot]$_{NP}$]$_{NP}$ remained [the residence of [the king]$_{NP}$]$_{NP}$ until [536]$_{NP \text{ but incompatible in semantics}}$ when [he]$_{NP}$ moved it$_{resolve}$ to [London]$_{NP}$.

Hobbs (1978) manually evaluated the algorithm on 300 pronouns from texts with gold parse trees. He found an accuracy of 88.3% without the additional semantic constraints and an accuracy of 91.7% when they were applied. Another evaluation of this algorithm (Tetreault, 2001) reports an accuracy of 76.4% on 1 694 pronouns from the *New York Times* subsection of the Penn Treebank Corpus (Marcus, Marcinkiewicz, and Santorini, 1993) and an accuracy of 80.1% on the 511 pronouns from the subsection of fictional texts. Tetreault (2001)'s evaluation was automatic but made use of the gold syntax trees from the corpora and manually-assigned gender features for the noun phrases in the texts.

Lappin and Leass' Algorithm

The algorithm of Lappin and Leass (1994) is an important reference for pronoun resolution. It is often referred to as (one of) the first robust algorithm(s) because it does not make use of world knowledge or discourse structure. It is a method that is based on four algorithms working together[1] to solve personal, reciprocal and reflexive pronouns. The first algorithm filters out antecedents that do not agree in gender and number or where coreference is ruled out by the syntactic structure of the sentence, such as in example (4).

(4) She$_i$ likes her$_{\neq i}$.

A second algorithm eliminates pleonastic (non referential) pronouns and a third algorithm is used to resolve reflexives an reciprocal pronouns. The fourth algorithm scores candidates according to their *saliency* that is based on a number of features. Interestingly, these features are very similar to the features that can be found in modern systems: grammatical function of the antecedent, syntactic parallelism, frequency of the antecedent and distance. If any of the antecedent candidates has a saliency score that surpasses a predefined threshold and has not be ruled out or resolved by the

[1] Note that the order of application of the four algorithms is not necessarily from one to four and depends on the type of pronoun.

other three algorithms, it is chosen as the referent of the pronoun. If more than one candidate surpass the threshold, the candidate with the best score is taken. Lappin and Leass (1994) report an accuracy of the algorithm on English of 86% when it was tested on 360 pronouns from a corpus containing instruction manuals for computers.

Centering Theory Based Algorithms

A third important reference for the early period of pronoun resolution are the Centering Theory (Grosz, Joshi, and Weinstein, 1983; Grosz, Weinstein, and Joshi, 1995) based algorithms. According to Centering Theory, there are two levels of discourse coherence: *global focusing* and *centering*. The first level is about the focusing on entities relevant to the global discourse and the second level refers to a more local focusing process related to *"identifying the single entity that an individual utterance most centrally concerns"*. These two levels of coherence are claimed to have different effects on the processing of pronouns and definite noun phrases.

According to Centering Theory, every utterance U of a discourse has a single backward-looking center C_b and multiple forward-looking centers C_f. Centers are *"the sort of objects that can serve as the semantic interpretations of singular noun phrases"*, for example, people, numbers, situations or objects (Grosz, Joshi, and Weinstein, 1983). The backward-looking center of an utterance U_{n+1} corresponds to one of the forward-looking centers of the previous utterance U_n. The elements of C_f are ordered by the relative prominence they have in U_n. The highest ranked forward looking center of U_n that is realized in U_{n+1} is the backward-looking center of U_{n+1}.

Centering Theory defines three types of relations that can occur between two utterances by which the coherence of a segment is affected:

(5) a. Center continuation: the backward-looking center of U_{n+1} is the same as the backwards-looking center of U_n and it is also the highest ranked element in the forward-looking center of U_{n+1}.
b. Center retaining: the backward-looking center of U_{n+1} is the same as the backwards-looking center of U_n, but this entity is not the highest ranked element in U_{n+1}'s forward-looking centers.
c. Center shifting: the C_b of U_{n+1} is different from the C_b of U_n.

Obviously, there need to be criteria to order the C_fs. Criteria that make C_fs more prominent and hence more likely to become the C_b are based on their syntactic function — where subjects are ranked higher than other grammatical roles — and surface position, where referents that are included first are given more prominence (Grosz, Weinstein, and Joshi, 1995).

In addition to the ordering criteria of the C_fs and the centering relations (5), Centering Theory gives two rules of "centering management":

(6) a. **Rule 1:** If any element of $C_f(U_n)$ is realized as a pronoun in U_{n+1}, then the $C_b(U_{n+1})$ must also be realized by a pronoun.
b. **Rule 2:** Sequences of continuation are preferred over sequences of retaining; and sequences of retaining are to be preferred over sequences of shifting.

Based on Centering Theory, algorithms of pronoun resolution have been developed. Brennan, Friedman, and Pollard (1987)'s is one of the best known. This algorithm

performs pronoun resolution by determining all possible C_bs and C_fs for a discourse, filtering out impossible solutions with regard to the centering rules and finding the pronoun resolutions that result in the best transitions, according to Rule 2 (6-b). Tetreault (2001) found an accuracy of 59.4% when he evaluated this algorithm on 1694 pronouns from the *New York Times* subsection of the Penn Treebank Corpus (Marcus, Marcinkiewicz, and Santorini, 1993) and an accuracy of 46.4% on the 511 pronouns from the subsection of fictional texts.

Knowledge Poor Approach

An important step in the transition towards fully automatic systems was made with the *knowledge poor* approach (Mitkov, 1998). The proposed system is quite similar to the saliency algorithm of Lappin and Leass (1994): several features are used to score antecedent candidates of an anaphoric pronoun. The candidate with the highest score is predicted to be the antecedent. The main difference between this work and the one of Lappin and Leass (1994), is that it does not need a parser. The features are very simple. According to Mitkov (1998), the accuracy of 89% they obtained by evaluating the algorithm on texts of technical nature after manual processing, is comparable that of Lappin and Leass (1994). However, both results are not fully comparable since they were not obtained on the same data.

3.2.2 Corpora

Since the 1990's, annotated corpora have played an increasingly important role in the field of Natural Language Processing. With this development, the first data-driven techniques appeared in the field of anaphora resolution and coreference resolution. The data-driven methods developed hand in hand with shared task data sets. Shared tasks allowed researchers to evaluate their systems on standardized data sets. As a result, systems became far more comparable. In this section, we discuss the shared tasks and their data sets. In Section 3.2.3, we discuss how the performance of resolution systems that participate in the shared tasks is evaluated.

Shared Tasks and Standardized Data Sets

The data sets that are the most used in research on coreference resolution for English are the MUC-corpora, the ACE corpora and the OntoNotes corpus. In this section we quickly review them. The goal of this section is that the reader can have an idea of the role these corpora play and understand how automatic systems are evaluated. Of course, there are other corpora that specialize in other languages, or in other genres — even though most corpora remain of the newspaper genre. We refer the reader who is interested in these resources to the work of Poesio et al. (2016), who provide a quite extensive overview of resources of anaphora and coreference resolution. We also refer to this work if the reader wants more details about what is considered as a mention in each resource.

The 6^{th} and the 7^{th} edition of the *Message Understanding Conference* (MUC-6 and MUC-7) in 1996 and 1998, laid the basis of the coreference resolution task (Poesio et al., 2016). For the shared tasks proposed at these conferences, the first corpora large enough to train and evaluate automatic systems were developed: the MUC-6 and MUC-7 data sets. The annotation schemes that were used form the foundation of

the schemes used today. They are rather simple: mentions, or *markables*, that belong to a coreference chain receive an identifier (*id*), the *id*-number of the first mention in the coreference chain. Both MUC-corpora contain about 30 000 words. All mentions in the MUC-corpora are nominal. Event relations, bridging relations and abstract relations are not annotated. But the predominance of nominal relations is typical of most coreference annotated corpora.

The ACE-corpora were used for the *Automatic Content Extraction program* from the year 2000 until 2008 (Poesio et al., 2016). They contain data for three languages: English, Chinese and Arabic. The English corpora replaced the MUC-corpora as a standard evaluation corpus.

Today, an important place is taken by the OntoNotes corpus (Pradhan et al., 2011). According to Poesio et al. (2016), this corpus is the largest coreference annotated corpus for English (1.6 million words), Chinese (1 million words) and Arabic (300 000 words). The corpus contains a diversity in genres and also contains a gold annotation of syntactic trees that makes it easier to detect mentions before passing them to humans for annotation. The inter-annotator agreement of this corpus is high and everything has been annotated by two annotators in a double-blind manner (Poesio et al., 2016). The corpus is freely available and the combinations of these features make it an important evaluation and training corpus for the NLP-community working on coreference. Later in this chapter, we will discuss this corpus in more detail, because we used it to train the pronoun resolution system we implemented for our experiments.

3.2.3 Evaluation Metrics

With the shared tasks, objective methods were also developed to evaluate the coreference systems. The evaluation of coreference resolution is similar to the evaluation of clustering (Luo and Pradhan, 2016). But there is an important difference. In a normal clustering problem, the elements before the clustering can all be retrieved after the clustering, the only difference is that they are regrouped together. A comparison between the system's output and the gold clustering is made to assess system performance. But, in the case of coreference resolution, mention detection is part of the problem. Before clustering mentions, it has to be decided what a mention is. It can therefore happen that the mentions in the gold partition in the corpus do not correspond completely to the mentions detected by the system (Luo and Pradhan, 2016). This makes the coreference resolution evaluation more complex than normal clustering evaluation.

Even though many more evaluation metrics have been proposed, four are key in the domain of coreference research: MUC, B^3, CEAF and BLANC. These four metrics have been described very well by Luo and Pradhan (2016) and for details about the maths and elaborated examples, we would like to refer the reader to this work. We will try to explain the general motivation and intuitions behind these four different metrics. In particular, we will explain how they penalize errors. Indeed, when mentions are regrouped into clusters there are many ways to do so. As a consequence, a multitude of evaluation metrics has been developed. Our presentation of evaluation metrics follows Luo and Pradhan (2016). We refer the reader to their review for details.

MUC

MUC was the first metric proposed for the evaluation of coreference resolution for the shared tasks of the Message Understanding Conference. It is an F-measure that relies on the recall and the precision of the links between mentions inside the entities of the system's partition (response partition) and those of the gold annotation (key partition). A link is a connection that is assumed between two subsequent mentions inside one coreference chain. Because coreference evaluation is based on the number of links inside an entity, the MUC-system is not adapted to evaluate singleton entities (entities with only one mention in them). The reason is that inside a singleton entity there are no links, so precision and recall cannot be defined for these entities. Moreover, spurious links (links predicted by the system, but absent in the gold standard), are not always penalized as we would like them to be. For example, when the gold partition is $\{a,b,c\},\{d,e,f\},\{g\}$, and the system's response is $\{a,b,c,d,e,f\},\{g\}$ this partition gets a better score than $\{a,b\},\{d,e,f\},\{g,c\}$, whereas our intuition would say that the second system's partition might be better.

B^3

B^3 is an F-score and was designed to correct the undesired properties of the MUC-score. The crucial difference between this metric and the MUC metric is that it is based on the number of mentions that entities in the key partition and the response partition have in common, instead of being based on the number of links in common. This solves the problem the MUC-metric has with singleton mentions, because even if there are no links inside singleton entities, they still contain one mention. But the B^3-metric leads to a new problem: in order for the B^3-metric to be valid, every mention can only occur once in the response. If mentions occur more than once, the recall of the B^3-metric may exceed 1 and err the metric. According to Luo and Pradhan (2016), it is not a borderline case to include mentions in more than one entity, but rather common to many coreference resolution systems.

CEAF

CEAF, which stands for *constrained entity-aligned F-measure*, was designed to overcome the problem of the B^3-metric. The CEAF-metric first finds an optimal alignment between the key partition and the response partition, before comparing the similarity between the mentions of the key and the response partition, or the similarity of the entities between these two partitions. Hence, the CEAF-metric can be applied on an entity-level as well as on a mention-level.

BLANC

BLANC is another alternative for the MUC-metric. BLANC stands for *bilateral assessment of noun-phrase coreference*. It considers both within entity links and cross-entity links. The strong point of the BLANC-metric is that it can give an F-score for coreference links but also for non-coreference links. Originally, it was only applicable when the response mentions were identical to the key mentions. This is not often the case for coreference resolution systems because the mention detection is part of the coreference resolution task. However, the BLANC-metric may be adapted to bypass this limitation (Luo and Pradhan, 2016).

3.2.4 Modern Systems

In this section, we discuss some key works and approaches from roughly the millennium change until now. The goal is to explain some influential algorithms, that are present in many systems as well as current debates and today's state of the art in coreference resolution.

Pair-wise algorithm

The mention-pair algorithm of Soon, Ng, and Lim (2001) is important because its architecture is very often reused in later works. Its simplicity makes it very attractive. The link with previous works, such as Lappin and Leass (1994) is also clearly visible: both works have many classification features in common.

The algorithm uses supervised machine-learning. Soon, Ng, and Lim (2001) used the decision tree algorithm, but other supervised machine-learning algorithms could also be used. The goal of the algorithm is to decide for two mentions whether they are coreferent or not. We could formulate this objective as Equation 3.2 in the case of a binomial outcome or as Equation 3.3 in the case of a probabilistic outcome.

$$coreference(m_i, m_j) \in [True, False] \tag{3.2}$$

$$P((m_i, m_j) = coreferent) \tag{3.3}$$

The algorithm is trained on examples of coreferent and non-coreferent pairs of mentions. In this way, it can learn representations of these two classes. The training examples are extracted from annotated corpora. Positive examples are obtained by putting two mentions from the same coreference chain into a pair. Negative examples are formed by pairing a mention with all the mentions that lie in between itself and its closest antecedent. The reason to restrict the negative examples in this way is to prevent an unbalanced set of training examples where there are too many negative examples.

As explained in Equations 3.2 and 3.3, the mention-pair algorithm is designed to decide whether two mentions are coreferent. However, the task of coreference resolution asks for coreference chains. Thus, in addition to the pair-wise algorithm, there must be a method to come to a final coreference clustering. Soon, Ng, and Lim (2001) solve this problem by forming chains incrementally. For every mention m_i in a document, it is decided whether it should be added to an already existing chain, or that it stands on its own. To decide whether the mention m_i should be added to an existing chain, pairs between m_i and all preceding mentions in the document (or in a restricted history of the document) are formed. So for all preceding mentions m_j with $j \in [0, 1, ..., i-1]$, the pairs $\langle (m_i, m_0), (m_i, m_1), ..., m_i, (m_{i-1},)\rangle$ are formed. When only one of the mention-pairs m_j with $j \in [0, 1, ..., i-1]$ is classified as positive (*true*, or $P((m_i, m_j) = coreferent) > 0.5$), m_i is attached to the m_j in question. If more than one mention-pair (m_i,m_j) are classified positive, a heuristic is used. There are two heuristics: *closest first* — introduced by Soon, Ng, and Lim (2001) — and *best first* — introduced in Ng and Cardie (2002). For *closest first*, m_i is attached to the m_j that is the closest in distance in the text. For *best first*, m_i is attached to the m_j that has the highest probability score $P((m_i, m_j) = coreferent)$. Note that for *best first*, a simple binary classifier cannot decide, as the only possible outcomes are *true* or *false*.

If none of the m_i, m_j pairs is classified as positive, the mention m_i forms a cluster on its own.

Soon, Ng, and Lim (2001) evaluated their algorithm on the test sets of MUC-6 and MUC-7 and reported respectively a MUC-score of 62.6 and 60.4.

Ranking Algorithm

An improvement on the pair-wise classification method is a ranking approach. Denis and Baldridge (2007) investigated how much pronoun resolution improved when a ranker was used instead of a pair-wise classifier. The difference between pair-wise classification and ranking is the following: in a classification system, for every mention-pair (m_i, m_j) it needs to be decided whether these mentions are coreferent. In a probabilistic classifier this is done by calculating $P((m_i, m_j) = \mathit{corefent})$. Only one candidate is considered at the time, that is to say that the score for the pair (m_i, m_j) does not take into account the score for the pair (m_i, m_{j-1}), even though this score is also about resolving the mention m_i. A ranking system, on the other hand, attributes scores to all antecedent candidates at the same time. As a result, a final clustering heuristic such as *closest first*, or *best first* becomes superfluous.

Denis and Baldridge (2007) show that the ranking method leads to a substantial improvement on the task of pronoun resolution compared to a pairwise approach. They evaluated the two approaches (and a third approach that is not relevant to this thesis), on the referential personal pronouns and possessive pronouns of the ACE corpus (from the second phase of the evaluation campaign). They reported an accuracy of 66.8% for finding the antecedent of an referential pronoun on their held out test set for the pair-wise approach and of 74% for the ranking approach.

The evaluation provided by Denis and Baldridge (2007) is important for our research: note that their accuracy seems lower than the one reported by Mitkov (1998), Hobbs (1978) and Lappin and Leass (1994). However, the results are not comparable because the way Denis and Baldridge (2007) evaluated the data does not involve manual preprocessing, they used more data and they evaluated on a corpus that is accessible for comparison. The high scores of Mitkov (1998) and Lappin and Leass (1994) could also be a result of testing on the text genre of technical manuals. Probably these documents are written in a least ambiguous manner. Therefore, we can consider that the score obtained by Denis and Baldridge (2007) is far more reliable and defines a better standard to compare against.

Mention-based versus Entity-based Approaches

With the higher interest in coreference resolution, the question whether resolution should take place on the entity, or the mention level, or both, has arisen. The coreference resolution task can be defined as putting mentions that refer to the same entity together.

Basically, such a partition can be obtained by the method of attaching mentions together and derive a cluster by the principle of transitivity. If mention A is attached to mention B and mention B attached to mention C, then, by the principle of transitivity mention A must be attached to mention C, leading to the cluster $\{A, B, C\}$. However, it should be clear that this method is sensitive to error propagation. Imagine that $\{A, B\}$ is clustered because of a high score for the pair (A, B) and $\{C, D\}$ because of a high score between mention C and mention D. If there is also a marginal

positive score between mention B and mention D — which is the result of an error — then the cluster $\{A, B, C, D\}$ emerges. This cluster is problematic because it assumes not only erroneously the link (B, D), but also the identity relations (A, D), (A, C) and (B, C).

A second strategy is not to attach two mentions together, but a mention to a cluster. This is intuitively the way that coreference resolution is thought of. Instead of building representations only over individual mentions and mention pairs, features for whole clusters are designed. However, it seems to be the case that it is more difficult to design effective features for clusters than for mentions. Cluster features must specify the set of mentions that the cluster contains. Therefore, the features often reflect a kind of a mean value for all mentions in the cluster, such as *most mentions in this clusters are of feminine gender* (Wiseman, Rush, and Shieber, 2016). Practices of this kind make the features less effective. Another way to obtain cluster features is to combine the feature values of the mentions in the cluster. If two noun phrases and one pronoun are in a cluster together, the category of the cluster could be represented as the value *NP-NP-PRO*. But the downside is that it leads to a lot of sparsity in the representations.

Results of the use of entity-level features are mixed. Some authors find a benefit of using them, whereas others do not find that they improve the resolution. The debate about the necessity of entity-level features is still ongoing. But we will see in the section about neural networks (Section 3.2.4) that the modern flexible neural network structures give a new twist: entity level features do not need to be hard-coded because neural networks can learn them implicitly (Wiseman, Rush, and Shieber, 2016).

The Odd Man Out: The Stanford Multi-Pass Sieve Coreference Resolver

Amongst all the machine-based approaches, there is still one particular system that works rule-based. Lee et al. (2011) developed a system that applies a series of rules according to their precision. This means that rules that provoke the fewest errors are applied before those that provoke more errors. In this way, error propagation is limited. For example, the first rules to be applied are of the type *string matching*: the strings of characters corresponding to two mentions (that are non pronominal) are compared. All mentions that show a high match are grouped together. The mentions that are easy to solve are grouped together in an early stage. This makes it easier for the system to group more difficult mentions — such as pronouns — at a later stage. The strategy has enabled Lee et al. (2011)'s system to be the state of the art for a while. It obtained better scores than machine learning systems at the 2011's CONLL shared task.[2] We believe that the system's architecture is useful to distinguish between different types of coreference. It enables it to handle the nature of different coreference relations more efficiently.

This system has been very popular for a long time. We think that its success was also due to its integration into the Stanford NLP-Pipeline (Manning et al., 2014). Many systems that have been proposed for coreference resolution do not include

[2]For the open track — which means that external sources may be used besides the data provided by the shared task organisers — they obtained the following scores: MUC: 61.0; B^3: 68.9; CEAF (based on entities): 45.0; BLANC: 74.0.

preprocessing steps such as mention-detection and syntactic parsing. The integration of the coreference into the pipeline made it easier for researchers to use the system for their own purposes.

Neural Networks at Work

The progress of performance of coreference resolution today is — like for many other NLP-tasks — largely due to the use of more powerful machine learning techniques, based on neural network architectures. The literature shows that today's state of the art systems are still very much inspired by previous works. The classification, ranking and clustering features remain largely the same and the mention-pair and entity approach can still be found in today's state of the art. The reason for the better results of neural networks is that the networks can learn more complex functions than 'classical' machine learning techniques (for example linear classifiers) and are therefore better able of modelling the data. But there is a second reason that neural networks improve the state of the art: neural networks provide flexible architectures that makes it easy to combine several systems into a single model and form a joint model. Neural models can thus dispense with a classical pipeline architecture that is prone to error propagation.

Some good illustrations are the works of Wiseman et al. (2015), Wiseman, Rush, and Shieber (2016), Clark and Manning (2016a) and Clark and Manning (2016b). Wiseman et al. (2015) for example, proposes a ranking system that ranks antecedent candidates of mentions. They illustrate that complicated feature combinations, as proposed in previous work using linear classifiers, is not necessary any more, because their convolutional neural network is able to extract the necessary feature-combinations on its own. Wiseman, Rush, and Shieber (2016) expand on this model by adding a second neural network structure that builds an entity-level representation on the basis of the hidden layers of the mention-ranking system. Clark and Manning (2016a) and Clark and Manning (2016b) present models use a mention-ranking model on the one hand, and a second model that decides whether two clusters must be fused or not on the other hand.

The current state of the art, the end to end resolution system of Lee et al. (2017), also integrates several neural networks in one architecture. They combine a mention ranking system together with a mention-detection system. In this way, mention detection and coreference resolution are performed simultaneously, which reduces error-propagation from the mention-detection phase. A comparison of the performance of the neural network systems can be found in Table 3.1.

TABLE 3.1: The performance of the neural network based algorithms on the CoNLL 2012 shared task data.

	MUC	B^3	CEAF
Lee et al. (2017)	77.2	66.1	62.6
Clark and Manning (2016b)	74.6	63.4	59.2
Clark and Manning (2016a)	74.2	63.0	58.7
Wiseman, Rush, and Shieber (2016)	73.4	61.5	57.7
Wiseman et al. (2015)	72.6	60.5	57.1

In conclusion, neural networks provide a very flexible architecture that makes it easy to combine several systems into a single model. They can define very complex functions and feature selection is done internally, without the need of manual feature selection. Therefore, they improve the state of the art in anaphora and coreference resolution remarkably. Nevertheless, it remains a challenge to interpret these architectures: it is difficult to know what exactly it is that the model learned about anaphora and coreference representations.

3.3 An Interpretable Model of Pronoun Resolution

In this section, we present the model we trained to study the bias for the subject (see Section 2.2.1) and the bias to resolve to an antecedent that is in a parallel syntactic position (see Section 2.2.2).

3.3.1 Pronouns

We only accounted for third person singular personal pronoun resolution in order to approach the psycholinguistic domain where pronoun resolution is most often restricted to these type of pronouns. The third person pronouns can be viewed as different from the first and the second as the latter are deictic rather than anaphorical, meaning that they are not resolved by the discourse context, but rather by the extra-linguistic context. We also excluded possessives because, strictly speaking, they are not pronouns. They are syntactically dependent on a noun phrase and are therefore likely to be processed in a different way.

We did not implement a pronoun-detection module but used the corpus part-of-speech tags. We only took pronouns that were anaphora, according to the corpus annotation. We did not handle cataphora, neither non-referential (pleonastic) occurrences of the pronoun *it*. The reason for which we did not implement mention detection was that our goal is not to develop the best possible pronoun resolver but rather to measure pronoun resolution biases on corpus. Hence, to adequately measure these biases, it is important to reduce noise coming from pronouns that are not anaphora.

3.3.2 Resolution Algorithm

We used a classifier that proceeds according to a probabilistic version of the pair-wise classification algorithm (see Section 3.2.4) but only for third person personal pronouns, as explained in Section 3.3.1. We used the method of Soon, Ng, and Lim (2001) to sample training examples: to get positive training examples (coreferent pairs), each pronoun is coupled to its closest antecedent. To get negative training examples, the pronoun forms a pair with every mention occurring between its closest antecedent and itself.

3.3.3 Machine Learning Algorithm

We chose to implement the pair-wise algorithm with a logistic regression classifier. We chose it for its straightforward interpretation of feature weights, indicating the influence of factors in pronoun resolution.

Logistic regression is a technique very similar to multiple regression discussed in Section 2.3.3. Logistic regression is a method to predict a categorical variable from continuous and categorical variables, whereas multiple regression is used to predict a continuous variable by using also continuous and categorical variables (Field, 2009). As shown in Section 2.3.3, the multiple regression model is characterized by Equation 3.4, in which the values of the different features of x are represented by the $x_n-variables$.

$$y_i = \alpha + \beta_1 x_{i_1} + \beta_2 x_{i_2} + \cdots + \beta_n x_{i_n} + \epsilon_i \tag{3.4}$$

In logistic regression, the value of the outcome variable y is not predicted, but the probability that an instance of x belongs to the category of y is estimated (Field, 2009). The simplest form of logistic regression, with one predictor variable is given by the following equation:

$$P(Y) = \frac{1}{e^{-(\alpha+\beta_1 x_{i_1}+\epsilon_i)}} \tag{3.5}$$

This formula generalizes easily to the case of multiple predictor variables.

$$P(Y) = \frac{1}{e^{-(\alpha+\beta_1 x_{i_1}+\beta_2 x_{i_2}+\cdots+\beta_n x_{i_n}+\epsilon_i)}} \tag{3.6}$$

Just as in the case of multiple regression, the coefficients of the model — the $\beta_n-variables$ in Equations 3.5 and 3.6 — are estimated on the basis of the best fit with the data points. This method is called *maximum likelihood estimation* (Field, 2009).

3.3.4 Corpus

For our experiment, we used the English newswire part of the OntoNotes 5.0 corpus (Pradhan et al., 2011). This corpus is annotated with coreference information, syntactic Penn Treebank style parsing (Marcus, Marcinkiewicz, and Santorini, 1993) and named entity information (Pradhan et al., 2011). The newswire genre approximated the psycholinguistic items the best among the available genres in OntoNotes. We divided the corpus that consisted of 792 texts into a training corpus of 476 texts, a development corpus of 158 texts and a test corpus of the same size.

A particularity of the corpus is that singleton mentions (referential expressions that are only mentioned once) are not annotated. We resolved this problem by simply considering as a singleton mention every maximal noun phrase that did not overlap with an annotated mention and that was not a pronoun. Moreover, since OntoNotes is not annotated for number nor gender, we had to add (automatically) an annotation for number and gender to the mentions in the corpus.

We used the resource produced by Bergsma and Lin (2006) about gender information, that provides counts of word forms occurring as respectively male, female and neutral gender on the web, to annotate the mentions in our corpus. More precisely, we took the three lists of unigrams (one for each gender) from the Stanford Core NLP Toolkit (Manning et al., 2014) — that was compiled from the resource of Bergsma and Lin (2006) — to annotate each token of a mention in our corpus with gender if it occurred in one of the lists. Then we propagated the gender of the syntactic head to the entire mention. Finding the head of a mention was done using a heuristic: the head is the last word of the mention, except if there is a prepositional

phrase inside the mention, in the latter case the head of the mention is the word before any prepositional phrase.

The number annotation was only done for tokens that were common nouns and proper names. In the tag set of the corpus, singular common nouns are tagged as *NN*, singular proper names as *NNP*, plural common nouns as *NNS* and plural proper names as *NNPS*. We used these tags to assign number to tokens. Then, we proceeded with the same *head heuristic* as for the gender feature to assign number to the entire mention.

3.3.5 Features

The aim of our model is to have interpretable features and not to have the best score on a pronoun resolution task. We proceeded in three steps to establish the features of our classifier. First, we defined a list of standard features for pronoun resolution — inspired by the coreference resolution literature (Soon, Ng, and Lim, 2001; Yang et al., 2004; Denis and Baldridge, 2007; Recasens and Hovy, 2009) — that we could retrieve in our corpus. We implemented a computer program that could estimate all these features for the pronouns in our corpus and used the development set to verify whether we implemented them correctly.

Among all the features, we made sure we included the features necessary to test the two biases investigated in our experiment. For the Subject Assignment Strategy, we used a feature that checks whether the antecedent candidate is in the subject position. We did this by checking whether the syntactic constituent of the antecedent candidate was underneath a subject node. We implemented the Parallel Function Strategy by a boolean feature of *syntactic path match* that states whether the antecedent candidate and the pronoun have the same path in the syntactic parse tree from the node where the mention is attached to the root of the tree. A simple illustration is given in Figure 3.1, where the syntactic paths of two mentions are given. The reason that we did not just check for the syntactic function was that — except for the subject function — they are not annotated in the corpus. Using the syntactic path match is thus a somewhat stricter implementation of syntactic parallelism than just checking whether both mentions have the same syntactic function.

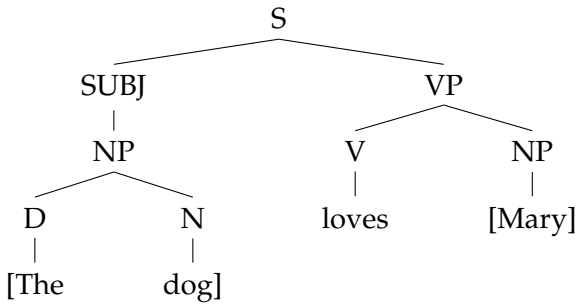


FIGURE 3.1: A syntactic tree with two mentions: *the dog* and *Mary*. Syntactic path for *the dog*: [SUBJ, S]. Syntactic path for *Mary*: [VP, S].

The second step of defining our features consisted in eliminating features too sparsely represented in our training corpus to be adequately learned. This concerned features that checked if mentions had a certain type of named entity, for example

EVENT or LAW, respectively occurring only 3 and 2 times in the training data. As a rule of the thumb we decided to exclude features with a frequency smaller than 0.5%, meaning that every feature should be attested at least 36 times in the training data. As a last step we checked the 95% confidence interval of our features' coefficients and removed features for which this interval contained 0. If the confidence interval contains 0, the interpretation of the features becomes difficult because we are not sure whether a higher value increases or diminishes the chance of coreference. A list of all features before model selection can be found in Table 3.2. We specify which features we keep for the final model and for the other features, we give the reason why they are eliminated. Our final model can be found in Table 3.3.

TABLE 3.2: The list of features we considered for our pronoun resolution model. m_1 is the potential antecedent, m_2 is the pronoun that needs to be resolved. Features are either kept for our final model or descarted because they are either too sparse or the confidence interval of their coefficient contains 0.

Feature	Decision
match in gender	keep
match in number	keep
m_1 is a subject	keep
match in syntactic path	keep
m_1 is a common noun	keep
m_1 is a proper name	keep
m_1 is a possessive pronoun	keep
m_1 is a personal pronoun	keep
mentions between m_1 and m_2	confidence interval contains 0
words between m_1 and m_2	keep
m_1 & m_2 in the same sentence	keep
length of syntactic path m_1	keep
m_1 is determined	keep
m_1 is undetermined	keep
m_1 has a demonstrative determiner	keep
m_1 spans m_2	keep
length of words of m_1	keep
number of occurrences of m_1 in the text	confidence interval contains 0
m_1 is a location	confidence interval contains 0
m_1 is a work of art	not enough data
m_1 is a geopolitical entity	keep
m_1 is an organization	not enough data
m_1 is a date	keep
m_1 is a product	not enough data
m_1 is a NORP[3]	not enough data
m_1 is a language	not enough data
m_1 is money	not enough data
m_1 is a person	confidence interval contains 0
m_1 is a law	not enough data
m_1 is an event	not enough data
m_1 is a quantity	not enough data

	Estimate	Signif.
(Intercept)	-2.3533	***
match in gender	2.4206	***
match in number	0.2430	*
m_1 is a subject	1.5142	***
match in syntactic path	1.7318	***
m_1 is a proper noun	0.5007	***
m_1 is a possessive pronoun	1.9037	***
m_1 is a personal pronoun	0.7647	***
words between m_1 and m_2	-0.0114	***
m_1 & m_2 in the same sentence	0.3587	***
length of syntactic path m_1	-0.1361	***
m_1 is determined	-0.2825	*
m_1 is undetermined	-0.4422	**
m_1 has a demonstrative determiner	0.6045	*
m_1 is a common noun	-0.8967	***
m_1 spans m_2	-3.4372	***
length in words of m_1	-0.0201	*
m_1 is a geopolitical entity	-1.2885	***
m_1 is a date	-1.9416	***

TABLE 3.3: The selected model of the pronoun resolver. Each factor influencing pronoun resolution has an estimated weight associated that indicates its coeficient. m_1 refers to the antecedent candidate, m_2 to the pronoun. Significance codes: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1.

3.3.6 Evaluation

We tested the model's performance on all three of the sets by measuring the accuracy of the identification of antecedents of the third person singular personal pronouns in the corpus. The accuracy and size for each subcorpus can be found in Table 3.4.

An important question is whether these results are satisfactory. Our results are difficult to compare against state-of-the-art work in coreference resolution, because we concentrate on third person personal singular pronouns only. This means that our system does not form coreference chains and that its performance cannot be measured using standard coreference evaluation metrics, such as MUC, B3, CEAF, or BLANC (Luo and Pradhan, 2016). A second difference with a more standard approach is that we do not have a module of mention detection. Instead, we use the gold mention annotation and the singleton mentions we extracted (see Section 3.3.4).

That said, we still want to have an indication about the performance of our classifier. One study that is similar to ours is by Yang et al. (2004), although they used a module for mention detection. Yang et al. (2004) trained different types of systems[4] to perform third person pronoun resolution and reported accuracy, in their paper indicated by the metric of *success*. When they tested on the MUC-6 corpus this metric was between 70.0 and 74.7 for the different systems they developed. When tested

[3] nationalities, organizations, religions, and political parties

[4] The systems differed in the features they used for training and the way training examples were constructed.

on the MUC-7 corpus the metric laid between 53.8 and 62.5. We estimate that, given these numbers, the performance of our model is slightly worse, or comparable. Another comparable study is by Denis and Baldridge (2007) cited in this chapter (see Section 3.2.4). They obtained a score of 66.8% on the ACE corpus for a mention-pair classifier designed to resolve pronouns and possessives. It seems to be the case that our classifier slightly underperforms their classifier, but again, the works are not comparable, because they are not evaluated on the same corpus and there is a difference in the type of anaphora that is resolved.

An error analysis we conducted on our system indicated that most of the errors made by the resolver concerned the pronoun *'it'* (about half of the errors). We observed that if we excluded *'it'* from resolution, the pronoun resolver's accuracy increased by $\approx$ 16 points. Our error analysis also indicated that a substantial proportion of the errors comes from our automatic gender annotation: it seems that many coreference chains contain mentions of several genders at once. Nevertheless, we think that the performance on masculine and feminine pronouns of our system is good enough for the purpose of our experiments on psycholinguistic materials that include only masculine and feminine pronouns.

Sub-Corpus	Nb. Texts	Nb. Pronouns	Accuracy
Training	476 (60%)	1756	61.15
Development	158 (20%)	558	65.41
Test	158 (20%)	617	61.26

TABLE 3.4: The accuracy of the resolver for finding the correct antecedent of the pronoun on the training, development and test set.

3.3.7 Interpretation of the Model

The weights of the logistic regression model in Table 3.3 predict the biases the classifier will show on experimental data. Looking at the feature of syntactic path match and the feature that checks if the first mention is in the subject position, we see that both features have a positive weight; but we can also see that the first is stronger than the second, suggesting that parallel roles are of a greater impact than the subject position of the antecedent. From this data we can hypothesize that the Subject Assignment Strategy exists alongside the Parallel Function Strategy, and that the Parallel Function Strategy, if applicable, has a stronger influence that can overrule the Subject Assignment Strategy.

3.4 Experiment: Simulation of Human Preferences

We saw that the coefficients estimated for the logistic regression model predicts that both the bias for the subject as well as a bias for an antecedent in a parallel syntactic function have an influence on pronoun resolution (see Section 3.3.7). To examine even more closely how plausible the biases learned by the model are, we propose to run the model on the experimental items that were used by Crawley, Stevenson, and Kleinman (1990) and Smyth (1994) to investigate them. Crawley, Stevenson, and Kleinman (1990) and Smyth (1994) both used experimental items that contained

ambiguous pronouns and asked participants to resolve them. The pronouns could either be resolved to an antecedent in the subject position, or to another one in an object position. This resulted in percentages of subject and object preference. In this experiment, we propose to test the model on these same experimental items and investigate the percentages of subject and object resolution. The purpose is to compare the model's percentages to the human percentages.

3.4.1 Procedure

We used the resolution system that we trained on the OntoNotes corpus, as presented in Section 3.3. We ran it on experimental items form the studies of Crawley, Stevenson, and Kleinman (1990) and (Smyth, 1994) in which participants were asked to resolve ambiguous pronouns. We also ran it on non-ambiguous items from Crawley, Stevenson, and Kleinman (1990)'s article. The reason was that we wanted to see how the model behaved on this non-ambiguous data. All the conditions we tested are explained and illustrated in Section 3.4.2.

Because our system uses syntactic features and named entity features, we annotated the experimental items. Before running the model, we manually annotated the items with coreference and named entity information. For the syntactic annotation, we first ran the Stanford Parser (Klein and Manning, 2003) and then manually corrected the parses.

We then compared the result of the model to the result of the human participants reported by Crawley, Stevenson, and Kleinman (1990) and Smyth (1994). Crawley, Stevenson, and Kleinman (1990) and Smyth (1994) reported the percentage of subject and object choices of participants per experimental condition. Crawley, Stevenson, and Kleinman (1990) had three experimental conditions: one ambiguous condition and two unambiguous conditions. In the ambiguous condition, the pronoun was an object and there was a parallel structure. In the two non-ambiguous conditions, the pronoun was also always the object, but in one condition the antecedent was the subject and in the other condition the antecedent was the object. Smyth (1994) used two ambiguous conditions in his experiment that we used in this study: one with the pronoun in the subject function and another with pronoun in the object function. We will illustrate all these conditions in more details in the Section 3.4.2.

3.4.2 Items

In this section, we present the items of the different experimental conditions used by Crawley, Stevenson, and Kleinman (1990) and Smyth (1994). For each type of item we give two examples to illustrate the type of experimental items used.

Crawley's Ambiguous Items

From the experiment of Crawley, Stevenson, and Kleinman (1990) we have 40 ambiguous items that can be found in the appendices of their article. Each item contains three sentences with a pronoun. In each item, two candidate mentions have the same gender and number, which makes the pronoun ambiguous. The pronoun that has to be resolved is presented in the last sentence in the direct or indirect object position. The two potential antecedents are equally in the direct or indirect object position. Crawley, Stevenson, and Kleinman (1990) designed the items to allow the parallel

function to operate. However, they did not make distinctions between direct and indirect object because in some items, such as the second example given here, display a potential antecedent in the direct object function and a pronoun in an indirect object function.

1. John and Sammy were playing in the garden. One of their classmates, Evelyn, tried to join in their game. John pushed Sammy and Evelyn kicked him.

2. Mary and Julie were about to go into town when they realized the car had a puncture. Their next door neighbour, Peter, was working in the garden. Mary helped Julie change the wheel and Peter talked to her.

Crawley's Unambiguous Items with Subject Antecedent

The unambiguous items had only one possible antecedent for the pronoun. The pronoun was always presented in the direct or indirect object position. In one of the two unambiguous conditions, the antecedent was in the subject function. The items are very similar to the items in the ambiguous condition of Crawley, Stevenson, and Kleinman (1990). Indeed, every item in the ambiguous condition has a counterpart in the unambiguous conditions. Note that in their article, Crawley, Stevenson, and Kleinman (1990) do not give a full list of the unambiguous items. Only an explanation how to obtain them from the ambiguous items is provided. Therefore, the items that are given as examples here are reconstructions we made following the instructions of Crawley, Stevenson, and Kleinman (1990).

1. John and Mary were playing in the garden. One of their classmates, Evelyn, tried to join in their game. John pushed Mary and Evelyn kicked him.

2. Mary and Tim were about to go into town when they realised the car had a puncture. Their next door neighbour, Peter, was working in the garden. Mary helped Tim change the wheel and Peter talked to her.

Crawley's Unambiguous Items with Object Antecedent

In the second unambiguous condition, the antecedent of the pronoun in the (in)direct object position, also always occurred in the (in)direct object position. Again, the 40 items in this condition are derived from the 40 ambiguous items. We also reconstructed these items based on Crawley, Stevenson, and Kleinman (1990)'s instructions.

1. Mary and John were playing in the garden. One of their classmates, Evelyn, tried to join in their game. Mary pushed John and Evelyn kicked him.

2. Tim and Mary were about to go into town when they realised the car had a puncture. Their next door neighbour, Peter, was working in the garden. Tim helped Mary change the wheel and Peter talked to her.

Smyth's Ambiguous Pronouns in Subject Position

We used 10 items from Smyth (1994) that were ambiguous items with a pronoun in the subject position. Parallelism is defined in a more strict sense than in Crawley,

Stevenson, and Kleinman (1990)'s items. First, Smyth (1994) distinguishes between the direct and the indirect object functions. Second, the items were shorter (only one sentence) and always consisted of two clauses that were connected by *and then*. Third, except the grammatical parallelism, the items also displayed lexical parallelism. For example, the same verbs were used in both of the clauses.

1. Mary helped Julie change the tire and then she helped Peter change the oil.
2. Shirley wrote to Carol about a meeting and then she wrote to Martin about a party.

Smyth's Ambiguous Pronouns in Object Position

Smyth (1994) also presents ten items with a pronoun in the direct or indirect object position. Except for this difference the items were constructed in the same manner as Smyth (1994)'s ambiguous items with pronouns in the subject position.

1. John pushed Sammy and then Evelyn kicked him.
2. Sarah visited Cathy at home and then Charles phoned her at work.

3.4.3 Results

We can see in Table 3.5 that the model fits human preferences quite accurately. With the ambiguous items from Crawley, Stevenson, and Kleinman (1990), we observed that the Subject Assignment Strategy applies as a default strategy when the Parallel Function Strategy is not available. For the unambiguous items, Crawley, Stevenson, and Kleinman (1990) did not report human assignment. The model's assignment for these items was 100% correct when the antecedent was a subject, but when it was an object or indirect object in 15% of the cases the model could not attribute a score high enough to choose it as the antecedent and responded *None*.[5] For the items of Smyth (1994)'s experiment, we observed — just like him — that the Parallel Function Strategy is the preferred strategy.

3.4.4 Discussion

We have shown that our model is able to mirror quite accurately pronoun resolution preferences. As our model is trained on real corpus data, we conclude that such preferences are somehow statistically present in the language. Our model is in line with the claim that the Parallel Function Strategy and the Subject Assignment Strategy exist alongside each other and that the former can overrule the latter. Our model embodies the idea Smyth (1994) has about pronoun resolution: *"Pronoun resolution is a feature-match process whereby the best antecedent is that which shares the most features with the pronoun."* It also captures Smyth (1994)'s idea that not every feature has the same impact and that for example *gender match* is more important than parallel roles. Based on the results our model obtained on the experimental items, we conclude that the weights of the subject preference and the parallel function preference it learned from corpus are cognitively plausible.

[5] Among all antecedent candidates the correct antecedent still got the highest score, but it was lower than 50%, so the resolver responded that it did not find the antecedent. This behaviour of the system can be seen as the result of training it on the OntoNotes corpus, where the bias towards classifying negative must be high, to prevent it from linking pronouns to wrong antecedents.

	Human		Model	
Experiment	% Sub.	% Obj.	% Sub.	% Obj.
Crawley, ambiguous items, pronoun in the object position	60%	40%	72.5%	27.5%
Crawley, unambiguous items, antecedent in the subject position	n.a.	n.a.	100%	0%
Crawley, unambiguous items, antecedent in the object position	n.a.	n.a.	0%	85%
Smyth exp. 2, ambiguous items, pronoun in the subject position	100%	0%	100%	0%
Smyth exp. 2, ambiguous items, pronoun in the object position	12%	88%	30%	70%

TABLE 3.5: Human pronoun assignment versus the model's predictions on Crawley, Stevenson, and Kleinman (1990)'s items and Smyth (1994)'s items from experiment 2. For each item set examples can be found in Section 3.4.2. For Crawley, Stevenson, and Kleinman (1990)'s unambiguous items, no human results were reported. Note that for the unambiguous items with pronouns in the object position, the model sometimes did not assign any antecedent to the pronoun.

3.5 Conclusion

In this chapter, we examined whether an automatic system could serve as a model of human pronoun resolution. We implemented an NLP-inspired resolution system and trained it on a large corpus annotated with coreference. When we looked at the model's coefficients, it seemed that resolution preferences described in the psycholinguistic literature were also reflected in the model. It is interesting that the model learned these biases from corpus data. It shows that the human-attested biases are also present as a general frequency tendency in the language.

The comparison of antecedent choice between humans and the model show that the choices of the model and human choices between subject and object antecedents for ambiguous pronouns are very similar. This is an argument in favour of the statement that NLP-models can make good cognitive models.

In the beginning of this chapter, we reviewed the NLP-literature about pronoun and coreference resolution. The NLP-community has done important work on the level of the scalability of the pronoun and coreference research by providing large corpora and establishing evaluation standards that enable researches to compare the quality of different systems.

The data-driven approach also launched debates about the problem of pronoun resolution, for example the role of the entity (cluster) as opposed to the role of the individual mention. To our knowledge, this debate does not play a role in psycholinguistic research, but the question whether human pronoun resolution and coreference resolution make use of entity representations or whether the resolution happens by linking two mentions, seems at least interesting. One of the reasons — according to us — that this issue has not been addressed yet is that it needs longer texts than those often used in psycholinguistic experiments.

A second discussion that we can find in the NLP-literature about coreference is the question whether different types of mentions — such as for example pronouns, full noun phrases or proper names — use the same resolution mechanisms. In psycholinguistics, there are some experiments that investigate the differences between the resolution of various types of referential expressions. For example, it has been pointed out that repeating a proper name leads to less successful processing than using a pronoun instead (Gordon, Grosz, and Gilliom, 1993). It was found that sentences such as (7-a) are processed more slowly than (7-b).

(7) a. Jason came home. Jason sat down on the couch.
b. Jason came home. He sat down on the couch.

However, in the field of psycholinguistics, the influence of referential expressions has not been studied yet on natural discourse. We therefore think that this would also be an interesting topic for future research and we believe that both the fields of psycholinguistics and NLP could help each other answering this question.

In conclusion, we can learn a lot from the models that have been proposed in the NLP-community and we could think of many ways to use the systems to build new cognitive models for pronoun resolution. In the next chapter, we will do so, exploring a state-of-the-art system to test a hypothesis we formulated about the cognitive load of pronoun resolution.

Chapter 4

Information Theoretic Cost Metrics

4.1 Introduction

Information Theory (Shannon and Weaver, 1949) is a mathematical theory of communication. It is an important theory to many domains of science. In the field of linguistics, it is used to measure information at various levels: from phonetics to syntax, or to estimate the efficiency human language in exchanging information. For example, notions from Information Theory can be used to evaluate the predictability of the morphological structure of languages (Beniamine, 2018).

Information Theory thus provides mathematical tools to evaluate the quantity of information that is transmitted through linguistic structures and the efficiency of linguistic communication. Cognitive hypotheses of language processing can be formulated with respect to information quantity and linguistic efficiency. For example, a cognitive hypothesis could be that more information leads to a higher cognitive cost of linguistic processing.

We give a brief presentation of information theoretical notions that have been used to model the processing of linguistic structures (Section 4.2). Then, in Section 4.3 we discuss the cognitive hypotheses based on these notions that have been formulated in the literature.

One of the goals of this thesis was to explore what information theoretical notions can bring to the topic of cognitive computational models of pronoun resolution. After a literature review about the studies that have proposed cognitive models of pronoun, anaphora and coreference resolution (Section 4.4), we present a hypothesis based on information theoretical notions about the cognitive load of pronoun processing.

At the end of this chapter, we present two experiments in which we tested our information theoretic cost metric for pronoun resolution. The first experiment, presented in Section 4.6, contains an evaluation of the hypothesis on reading time recorded for the items of the study of Crawley, Stevenson, and Kleinman (1990) (see Section 3.4.2 for more details about these items). We find that our information theoretic cost metric is able to simulate some but not all effects found in Crawley, Stevenson, and Kleinman (1990)'s study. In the second experiment presented in Section 4.7, our hypothesis is evaluated on the anaphoric pronouns of the Dundee Corpus (see Section 2.4 for a presentation of this resource). We find that our cost metric is a predictor of reading time: a result in line with our processing hypothesis.

4.2 Information Theory

In his paper entitled *A Mathematical Theory of Communication*, Shannon (1948) laid the foundation of what is now known under the name of Information Theory. The theory deals with the question of how a message can be encoded, sent and decoded in the most efficient way. It proposes to measure the information carried by messages by the number of bits that is needed to encode them.

When designing a communication channel, it is useful to know how much capacity it will need. Information Theory estimates the channel capacity by providing tools to calculate the number of bits that is expected given a certain type of messages. For example, if messages consist of random combinations of the letters $\{A, B, C\}$, the probabilities of finding A, B or C in the message determine how many bits will be needed on average to encode the messages.

Information Theory provides many areas of science with important tools to measure information and to estimate the efficiency of processes. In linguistics, these tools are also useful in order to measure the information and efficiency in linguistic communication. In the field of cognitive modelling, Information Theory is used to measure the amount of information and incertitude, upon which hypotheses can be formulated, for example: more information leads to more cognitive load.

In the following sections (4.2.1, 4.2.2, 4.2.3 and 4.2.4), we will explain the measures that are relevant for cognitive computational modelling. In Section 4.3, we discuss how these concepts serve to formulate hypotheses about linguistic processing.

4.2.1 Surprisal

Surprisal is also referred to self-information of an event. It is inversely related to its probability.

$$\text{Surprisal}(p) = -\log_2(p) \tag{4.1}$$

This means that the surprisal of an event with high probability is lower than that of an event with low probability (Sheldon, 1998). Intuitively we can explain it as follows: surprisal is greater for unexpected events than for expected events.

4.2.2 Entropy

Entropy captures how much 'uncertainty' plays a role in a given random variable. It corresponds to the number of bits that are needed on average to encode the outcome of a random variable (Thomas and Cover, 2006). Entropy is maximal if all the possible outcomes of the random variable have equal probabilities (uniform distribution). High uncertainty, or — in the cases of language processing — high ambiguity, gives a high entropy. In the formula of entropy, we see that the entropy is the sum of the surprisal of all possible outcomes of the random variable, weighted by the probabilities.

$$H(X) = -\sum_i p(i) \cdot \log_2(p(i)) \tag{4.2}$$

4.2.3 Relative Entropy

The notion of relative entropy, also referred to as the Kullback-Leibler divergence, captures how much dissimilarity there is between two probability distributions P and Q. Often, it is viewed as a distance between P and Q. However, this is actually incorrect because the relative entropy between P and Q is not the same as between Q and P (Thomas and Cover, 2006). The relative entropy has the following equation:

$$H_{relative}(P||Q) = \sum_{i \in P \wedge i \in Q} P(i) \log \frac{P(i)}{Q(i)} \tag{4.3}$$

4.2.4 Normalized Entropy

In this thesis, we use the term normalised entropy to design an entropy that is normalised, or scaled, by the maximal entropy. Regardless the number of possible outcomes, the normalised entropy will thus take values from the interval $[0, 1]$. In its formula, we recognize the formulation of simple entropy, divided by $\log_2 n$ — the maximal entropy for a variable with n outcomes:

$$H_{normalized}(X) = -\sum_{i \in X} \frac{p(X = i) \cdot \log_2(p(X = i))}{\log_2 n} \tag{4.4}$$

4.3 Information Theoretical Cost Metrics

In this section we discuss theories that use information theory to predict processing cost. We discuss Surprisal Theory (Hale, 2001), the Entropy Reduction Hypothesis (Hale, 2003; Hale, 2006) and the hypothesis of Uniform Information Density (Jaeger, 2010). These works were an important source of inspiration for the formulation of our own hypothesis about pronoun resolution presented in Section 4.5.

4.3.1 Surprisal Theory

Hale (2001) proposed a measure of cognitive load of language processing based on the notion of surprisal. Surprisal Theory hypothesises that the cognitive load for interpreting a word is dependent on how much the preceding context 'predicts' this word: highly predictable words are easier to process than unpredictable words. The predictability of a word can be estimated with a language model. A language model gives the probability that a word appears, given a context, or *history* (Martin and Jurafsky, 2009). The probability of a word given the history is formalized by the probability $P(w_n|h)$, where w_n is the n_{th} word and h the history. If this probability is estimated from a corpus, we can simply count the number of times that w_n occurs after the history h and divide it by the total number of counts of history h (see Equation 4.5).

$$P(w_n|h) = \frac{C(h + w_n)}{C(h)} \tag{4.5}$$

The history can be modelled in various ways. In theory, it could be modelled as all the words that appear previously in the text, which would come to assuming that $h = w_1^{n-1}$. However, when w_n is further in the text the long history would be

more and more difficult to find in a corpus. Hence, $P(w_n|h)$ becomes impossible to estimate because the history is not attested in the corpus. A solution is to assume that a shorter history is a good enough approximation. We could for example think of a very simple bi-gram model in which the history is reduced to the previous word (see Equation 4.6).

$$P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-1}) \tag{4.6}$$

In a bi-gram language model, for every word, a probability distribution is formed based on the words that follows. The probabilities are based on corpus counts. To take a very simple example, imagine that we took the following discourse as a corpus and were interested in the words 'Smurf' (in red) and 'each' (in blue):[1]

(1) The Smurfs' community generally takes the form of a cooperative, sharing, and kind environment based on the principle that each Smurf has something he or she is good at, and thus contributes it to Smurf society as he or she can. In return, each Smurf appears to be given their necessities of life, from housing and clothes to food without using any money in exchange.

If we were interested in the probability distribution of the words following 'Smurf', we would find that it occurs three times and therefore obtain the following:

$$P(\text{has}|\text{Smurf}) = \frac{1}{3}$$
$$P(\text{society}|\text{Smurf}) = \frac{1}{3}$$
$$P(\text{appears}|\text{Smurf}) = \frac{1}{3}.$$

On the other hand, if we were interested in the context of 'each', according to this corpus, it appears two times and is followed twice by 'Smurf'. Therefore, we would obtain:

$$P(\text{Smurf}|\text{each}) = \frac{2}{2}$$

Imagine that we would continue to read this Smurf text. If the word 'each' appeared again in this text, we would not be surprised if it would be followed by the word 'Smurf'. On the other hand, if the word 'Smurf' appeared, it would be more difficult to guess the following word.

Besides this very simple bi-gram model, one can imagine more complex models, not only n-gram models, but also syntactic models. When a syntactic model is used, the question is not how surprising a word is given some previous words, but how surprising a word is, given the syntactic analysis for the preceding words of the sentence. The surprisal of the word can be measured by either looking at the exact word form — which gives a measure called *lexical surprisal* — or by taking the surprisal of the word's part of speech category — which gives *part of speech surprisal*. Hale (2001) proposes to study the part of speech surprisal by making use of

[1] This discourse was taken from: https://en.wikipedia.org/wiki/The_Smurfs#Smurf_economy.

an incremental parser that processes the sentences of a text word by word (Stolcke, 1995).

The probabilities of the parser's rules are used to calculate the surprisal. These probabilities can, just like the bi-gram probabilities, be estimated on a corpus containing syntactic structures. For example, if 70% of the noun phrases in the corpus is expressed by a determiner followed by a noun and 30% by a proper name, the probabilities of the rules 2. and 3. in Table 4.1 are accurate.

TABLE 4.1: A probabilistic context free toy grammar

	Rule			probability
1.	S	$\rightarrow$	$NP\ VP$	1.0
2.	NP	$\rightarrow$	$Det\ N$	0.7
3.	NP	$\rightarrow$	$ProperName$	0.3
4.	VP	$\rightarrow$	V	1.0

Syntactic surprisal can be thought of in the following way: the parser treats the text word by word. For each word, the parser emits one or multiple hypotheses about which rules could derive the input. For example, if a sentence started with the word *The* and if we used the toy grammar in Table 4.2, the parser activated rule 1. , 2. and 3. because they are all compatible with the word *The*. The grammar in Table 4.2 states that *The* necessarily leads to an NP, but it is still unclear whether the NP will be derived by rule 2. or 3. of this grammar. If the second word the parser reads is *nice*, it becomes clear that only rule 3. can account for the input and that rule 2. has to be dismissed. The surprisal when reading the word *nice* is calculated with the probability of rule 3. : $-\log_2(0.3) = 1.74$. The surprisal of each word is thus equal to the surprisal of the total probability mass of the maintained rules. Surprisal is thus higher when less frequent structures are encountered. According to Hale (2001), surprisal is a measure of cognitive load. Surprisal Theory states that human processing cost is proportional to surprisal (Hale, 2001).

TABLE 4.2: A toy example of an extract from a probabilistic context free grammar.

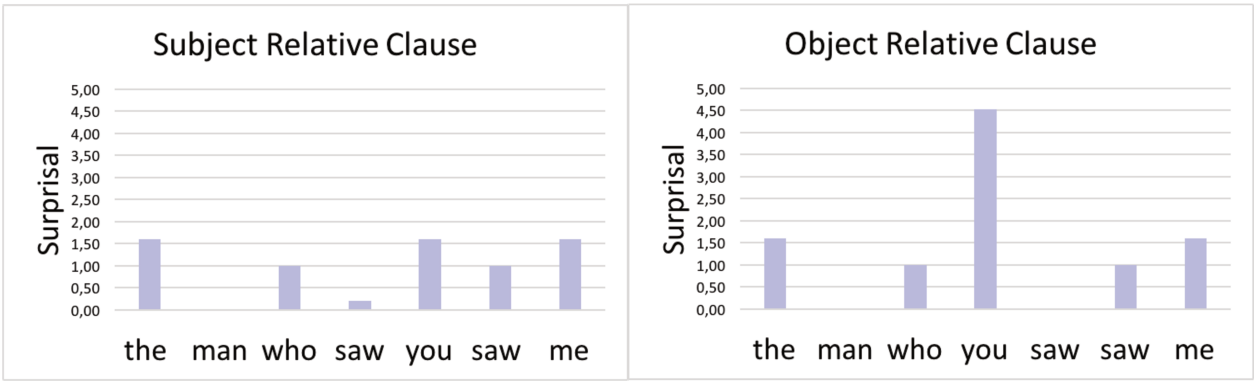
	Rule			probability
1.	S	$\rightarrow$	$NP\ VP$	1.0
2.	NP	$\rightarrow$	$Det\ N$	0.7
3.	NP	$\rightarrow$	$Det\ Adj\ N$	0.3
4.	VP	$\rightarrow$	V	1.0

Let us look at a more concrete example. In Hale (2001), surprisal for a subject relative clause and an object relative clause are calculated for each word. In subject relative clauses, the noun to which the relative clause is attached has the subject function in the relative clause. An example is Sentence (2-a). In object relative clauses, the noun to which the relative clause is attached is the object in the relative clause, as is the case in Sentence (2-b).

(2) a. The man who saw you saw me.
b. The man who you saw saw me.

In Figure 4.1, we can see that the model predicts more processing cost for the object relative clause than for the subject relative clause because the probability of the rule for the subject relative clause, $NP[+R]VP$, is higher than the rule for the object relative clause: $NP[+R]S/NP$. This prediction is in line with the fact that humans show more processing cost for object relative structure in English.

FIGURE 4.1: Surprisal for the words of a subject relative clause and an object relative clause based on a probabilistic context free grammar. The difference in likelihood between the rules in red that generate subject and object relative clauses, is at the root of the difference in surprisal for the word you. This figure was taken from Hale (2001).



Parsing Rule			Probability
NP	→	SPECNP NBAR	0.33
NP	→	*you*	0.33
NP	→	*me*	0.33
SPECNP	→	DT	1.0
NBAR	→	NBAR S[+R]	0.5
NBAR	→	N	0.5
S	→	NP VP	1.0
S[R+]	→	NP[+R] VP	0.869
S[R+]	→	NP[+R] S/NP	0.131
S/NP	→	NP VP/NP	1.0
VP/NP	→	V NP/NP	1.0
VP	→	V NP	1.0
V	→	*saw*	1.0
NP[+R]	→	*who*	1.0
DT	→	*the*	1.0
N	→	*man*	1.0
NP/NP	→	ϵ	1.0

An important work on surprisal theory and testing processing hypotheses on corpus data is the study of Demberg and Keller (2008). Demberg and Keller (2008) evaluated two theories of processing on corpus: the Dependency Locality Theory[2] (DLT, Gibson, 2000) and Surprisal Theory (Hale, 2001). They evaluated whether these two could predict reading times of the Dundee Corpus (Kennedy, Hill, and Pynte, 2003). They tested two versions of surprisal using a syntactic parser: one version in which surprisal was estimated for part-of-speech tags and one for lexicalized parser rules. They found that part-of-speech surprisal was a significant predictor of first fixation, first pass and total reading times. They did not find an effect of lexicalized surprisal. The reason was that among the control factors of the model, simple n-gram probabilities were included. These probabilities are arguably quite similar to lexical surprisal. Demberg and Keller (2008) also found some effects of DLT: whereas it was not a significant predictor when applied to all words of the corpus, it was a significant predictor of reading times of (a subset of) nouns and verbs. Interestingly, the DLT-integration cost was not much correlated to surprisal measures. This led to the conclusion that a theory of processing should both contain a measure of predictiveness and of memory load.

Another remarkable work on surprisal is Frank and Bod (2011). In a search of what would be an accurate representation of grammar and parsing in humans, Frank and Bod (2011) compare the fit of different models containing different measures of surprisal on the Dundee Corpus. The models only differ in the way they estimate the surprisal. Recall that surprisal can be estimated with a simple n-gram model, or on the rewriting rules of a parser. Frank and Bod (2011) looked which model got the best fit on the data and hypothesised that the model with the best fit would be the most accurate representation of human parsing. They found that simple non-syntactic language models performed better than syntactic grammars and hypothesized that therefore, it could be that humans do not build full syntactic structure for sentences. However, we want to draw the reader's attention to the fact that the neural network models that learned the non-syntactic language models were probably more powerful than the syntactic grammars used for the experiment. We thus think that this can be an alternative explanation for the better fit of the non-syntactic language models to the human data.

Summing up, surprisal theory has the advantage of being flexible in the model that is chosen to estimate probabilities. This also makes it easily extensible to other domains. Various attempts have been undertaken to use surprisal to formulate processing hypotheses about other linguistic domains than syntax. For example the work of Mitchell et al. (2010) exploits an integrated measure of surprisal and compositional semantics and Dubey, Keller, and Sturt (2013), which we will discuss more in Section 4.4.1, exploited an integrated measure of syntax, pragmatics and discourse.

4.3.2 Entropy Reduction Hypothesis

Based on the concept of entropy (see Section 4.2.2), Hale (2003) formulated the *Entropy Reduction Hypothesis*. According to the Entropy Reduction Hypothesis, a drop

[2] According to this theory, an integration cost can be calculated for discourse referents (nouns and verbs). The integration cost of a discourse referent *rf* depends on two factors. On the one hand, it is assessed whether *rf* is discourse new or discourse old; the former resulting in more processing cost than the latter. On the other hand, the integration cost of *rf* is determined by the number of intervening discourse referents between itself and its syntactic head: more referents lead to higher processing cost.

in entropy marks a point on which disambiguation is necessary (Hale, 2003, Hale, 2006). The entropy reduction hypothesis states that disambiguation demands processing cost. Therefore entropy reduction should be positively correlated to processing cost.

Like in his work on surprisal (Hale, 2001), Hale (2003) proposes to use the probability distribution on rewriting rules of a context free probabilistic grammar. The probability distribution that is used to calculate entropy is formed over complete parses of the sentence. Just like with surprisal theory, the sentences are processed word by word. For every incoming word, every parse that was compatible with the input so far is checked: is it still compatible with the incoming word, or should it be dismissed? Entropy before and after the dismiss of parses is compared. The difference between the two is what is called the *entropy reduction*. To take the notation of Linzen and Jaeger (2014), we define A^i as the set of all possible parses from the beginning of the sentence up to the word w_i. The entropy of the sentence after having integrated the word w_i is given by Equation 4.7.

$$H_{w_i} = -\sum_{a \in A_i} P(a) \log_2(P(a)) \tag{4.7}$$

The entropy reduction is then formulated as Equation 4.8.

$$ER = -\max\{H_{w_{n-1}} - H_{w_n}, 0\} \tag{4.8}$$

A challenge in the calculation of entropy reduction is recursive grammars. They allow an infinite number of derivations and therefore it is difficult to get a finite set of possible parses, which leads to difficulty in establishing a probability distribution. Hale (2003) uses two solutions for this problem. The first is to transform the grammar if it displays left recursion to an equivalent grammar with no left recursion, because it is this type of recursion that leads to an infinite number of derivations. The second is to use Grenander (1967)'s theorem to calculate the expected entropy of the daughters of non-terminal symbols on the right hand side of the parsing rules.

Linzen and Jaeger (2014) evaluated the Entropy Reduction Hypothesis. Interestingly, along the Entropy Reduction Hypothesis, an alternative hypothesis, the *Competition Hypothesis* (McRae, Spivey-Knowlton, and Tanenhaus, 1998; Tabor and Tanenhaus, 1999), was tested as well. The Competition Hypothesis predicts that processing cost should occur at the time that many hypotheses still apply, i.e before disambiguation. It states that when ambiguity is diminished after disambiguation , there is less processing cost. However, Linzen and Jaeger (2014) did not find evidence for the Competition Hypothesis and did find evidence supporting the Entropy Reduction Hypothesis.

Difference between Surprisal and Entropy Reduction

A little debate goes on in the community about what is exactly the difference between surprisal and entropy reduction. First of all, an obvious difference that Linzen and Jaeger (2014) point out: *"Surprisal predicts that the distribution over competing predicted elements should not affect reading times: if the conditional probability of a word A is $P(A|C)$, reading times on the word will be proportional to $-\log_2 P(A|C)$, regardless of whether the remaining probability mass is distributed among two or a hundred options."*

That means that surprisal is only affected by the probability of one event of the possible outcomes of the random variable, whereas entropy is effected by the probability distribution as a whole. However, it should be clear though that in many cases there can be a large entropy reduction and also a large surprisal. Therefore, the two factors can be easily confounded.

A study that tried to distinguish the two is described in the article of Frank (2013). He found positive evidence for both surprisal and entropy reduction in a self-paced reading experiment. He found that some variance in the data is explained by entropy reduction alone and is independent from surprisal. He states that it is tempting to conclude that the two metrics represent distinct cognitive mechanisms. But his data only showed that participants who have a larger effect of surprisal have a larger effect of entropy reduction. He therefore concludes that the question whether surprisal and entropy reflect different cognitive mechanisms of language processing is still an open question.

4.3.3 Uniform Information Density

The *Uniform Information Density Hypothesis* was proposed by Jaeger (2010). It states that language is organised in such a way that the amount of information that is communicated stays stable at every time point of an utterance. It is based on the assumption from Information Theory (Shannon and Weaver, 1949) that there is a communication channel (noisy channel) with a given capacity (bandwidth). The most efficient messages constantly contain close to the maximal amount of information that would fit in the channel: not much more and not much less. According to Jaeger, this should be valid for any linguistic unit. So, speakers spread information equally across the utterance no matter whether we are looking at phonetics, syntax, semantics or discourse. To take an example: there is a correlation that shorter words (containing less phonetic information) are more predictable than longer words (Jaeger, 2010) and Piantadosi, Tily, and Gibson (2011) show that *information* — a measure they define in a way very similar to surprisal (see Section 4.2.1) — is a very good predictor of word length.

There are a lot of cases in which a speaker can choose to use a long form, or a short form (Jaeger, 2010). An example is the optional use of the complementizer *that* as illustrated in Example (3).

(3) This is the friend (that) I told you about.

In his article, Jaeger (2010) provides evidence from corpora that the complementizer *that* is guided by the principle of Uniform Information Density.

Note however that the Uniform Information Density Hypothesis only makes claims about the production of language. As our research is focussed on comprehension, this hypothesis is somewhat less relevant for our research than the Surprisal Theory or the Entropy Reduction Hypothesis. But, even if this work is not directly applicable on our research, we found it interesting to mention, because it also makes claims about anaphora.

For example in the work of Kravtchenko (2014), it is investigated whether more predictable referents are referred to with shorter referential expressions in Russian. She found that the predictability of the referents plays a role in the use of null subjects in Russian. When the referent is more predictable, there is more chance that a null-subject is used, supporting the Uniform Information Density Hypothesis.

4.4 State of the Art in Anaphora Resolution

In this section, we review the computational cognitive models of anaphora and coreference resolution that have been proposed in the literature. We are aware of three proposals that investigate cost metrics of anaphora and or coreference resolution on corpus data. We give this overview to the reader before explaining our own proposal, so that it becomes clear how our work fits into the literature landscape. It is important to note that these proposals are rather independent from each other. Also, our own information theoretical cost metric that is presented in Section 4.5 is not directly a continuation or improvement on one of these models, even though our research has features in common with the works presented in Section 4.4.1 and Section 4.4.3 that it is inspired by Information Theory and uses the Dundee Corpus. The purpose of this section is thus to give an overview of the few studies that had objectives similar to ours.

4.4.1 Discourse Surprisal

An interesting cost metric based on the notion of surprisal is proposed in the work of Dubey, Keller, and Sturt (2013). Dubey, Keller, and Sturt (2013) state that processing load of linguistic structures is caused by multiple linguistic factors at once. They invented an integrated measure of surprisal that, according to the authors, can capture surprisal from discourse phenomena in addition to syntactic surprisal. Their hypothesis is that new discourse entities are harder to process than discourse old entities.

A cost metric was formulated on the basis of syntactic surprisal as described in Section 4.3.1. Dubey, Keller, and Sturt (2013) estimated probabilities for syntactic structure and anaphoricity in a simple way: a Hidden Markov Model was implemented to chunk the sentences of the discourse and decide on anaphoricity at the same time. Gender and number features together with string-match were used to determined anaphoricity. If the tool said the mention was a new discourse entity, the surprisal of the syntactic structure is incremented. This means that the model predicts that new discourse entities demand more processing cost than old discourse entities.

The authors tested their cost metric by comparing the fit of a model with only syntactic surprisal to the fit of the model with the integrated measure on the prediction of reading time in the Dundee eye-tracking corpus (Kennedy, Hill, and Pynte, 2003). The model with the integrated measure had a significantly better fit.

Dubey, Keller, and Sturt (2013)'s study shows that coreference has an influence on reading times. Processing cost is higher when new discourse referents appear in the text and need to be integrated. This makes Dubey, Keller, and Sturt (2013)'s study the first work on a large uncontrolled corpus of natural text reading that showed this. It shows that there is potentially a lot more to investigate about this topic. The model is based on only one very simple feature: is a mention discourse old or discourse new? We believe that it is not plausible that this one feature alone can capture the entire influence coreference resolution has.

4.4.2 Reasoning Based on World Knowledge

The work of Frank et al. (2007) describes a model of pronoun resolution that, in contrast to Dubey, Keller, and Sturt (2013), includes many features that influence the model's prediction of the cognitive load of resolution. Frank et al. (2007) proposed a computational model that predicts resolution processing cost. They verified the model using reading time data from experiments in which participants had to resolve ambiguous pronouns.

The model they built is based on a distributed situation space model. This model makes inferences on fourteen events of a micro-world. An example of such an event is *The sun shines*, or *Bob is outside*. These fourteen events are represented by propositional logic and can be combined with each other with the logical connectors *and* and *or*. Sentences are represented by high-dimensional vectors. The dimensions of the vectors correspond to co-occurrence constraints that were learned on a hand-constructing training set of stories. An example of a story composed of three vectors is given in Example (4).

(4) a. The sun shines.
b. Bob and Joe play soccer.
c. Joe wins.

When a story is processed, the sentence vectors are integrated incrementally into one situation representation. The stories can present an ambiguous pronoun. This pronoun can either refer to *Bob*, or *Joe*, because these are the only characters present among the fourteen events of the micro-world. An example of an ambiguous pronoun in a discourse is:

(5) Bob is tired and Joe is not, so he wins.

When an ambiguous pronoun is encountered in a discourse, a representation for both the situations *Bob wins* and *Joe wins* is constructed. These two representations are called *attractor regions* and both 'pull' on the vector containing the ambiguous statement. When one of the attractor regions pulls hard enough, the vector 'falls' on this attractor region. This means that the ambiguity is resolved to this interpretation.

The pulling-mechanism is driven by a mechanism that updates vectors. This updating process determines the speed of the resolution. The speed of resolution is used to model reading times of humans. Frank et al. (2007) shows that the model can replicate human behaviour in sentences in which the first mention bias (see Section 2.2.3) and implicit causality[3] play a role.

[3]Implicit causality is a property some verbs have to draw attention on one of their syntactic arguments. For example in the sentence (i-a), special attention is drawn on Anna, the subject, and in (i-b) on *Mary*, the object.

(i) a. Anna amazed Peter.
b. Boris disliked Mary.

Attention is drawn to the argument that is the most likely to provide an answer to the question "Why?": in (i-a) we ask ourselves what is it that is so amazing about Anna and in (i-b) we ask ourselves what is so dislikeable about Mary. According to Bott and Solstad (2014), if it is possible to continue a sentence including an implicit causality verb with a *because clause*, doing so becomes the default processing strategy.

The model of Frank et al. (2007) is interesting, because it captures an inference process and links it to processing cost. Also the combination of resolution preferences and reasoning makes the work relevant. But it is clear that the model is very difficult to scale up: the model is already very complex, but only contains fourteen events. It would therefore be difficult to use this type of approach in the modelling of natural data.

4.4.3 Coreference and Focus in Reading Times

A recent study has examined the influence of *focus effects* on coreference resolution (Jaffe, Shain, and Schuler, 2018). The process of *focussing* is defined by Jaffe, Shain, and Schuler (2018) in the following way: *"Linguistic focus directs subjects' attention toward particularly salient or important discourse referents during sentence processing."*. An example is the use of an it-cleft in a sentence like (6) on *malaria*.

(6) It was malaria$_{focus}$ that Sophia got infected with in Tanzania.

According to Jaffe, Shain, and Schuler (2018), focussing plays an important role in coreference resolution: mentions of referents that are focussed are processed more quickly. They evaluated this hypothesis by testing two approximations of focus on a reading time corpus. The corpus they used was the Natural Story Corpus (Futrell et al., 2018). They annotated it for coreference using basically the same annotation scheme as the OntoNotes Corpus (Pradhan et al., 2011).

The two approximations for focussing that are proposed are *frequency-based* and *recency-based*. For every mention in the corpus, the frequency-based approximation counts the number of times that the entity that a mention refers to was used in the text. So, for a mention that is not coreferent, the value of this feature would be 0, for entities that were mentioned once before 1, etc. The mention count variable should test whether more frequent referents are easier to retrieve from memory. The recency-based approximation of focus is represented by two variables: the number of words between the mention and its referent (this variable is set to 0 for first mentions) and the number of intervening mentions. The assumption that underlies the mention-recency variable is that more recently mentioned entities are more salient and can be retrieved easier from memory.

The Natural Story Corpus (Futrell et al., 2018) is a resource in between data from classical controlled psycholinguistic experiments and natural text data, such as the Dundee Corpus (Kennedy, Hill, and Pynte, 2003). The corpus contains stories that sound natural, but that have been manipulated to contain low frequency phenomena of language. This enables linguists to study low-frequency phenomena in a natural setting. The corpus comes with reading time data from 181 participants that performed a self-paced reading task[4] on it.

Jaffe, Shain, and Schuler (2018) studied the influence of their focus approximation factors (frequency-based and recency-based) using linear mixed effects. The approach is similar to the one presented in Chapter 2: control factors, such as measurements of surprisal, were used in the model and then the variables that approximated focus. An important difference with the work from Chapter 2 is that the number of participants was sufficient to estimate a random intercept and random slope for all the variables of the model: the control factors and the focus factors. A

[4]See Footnote 7 in Section 2.3.3 for an explanation about the self-paced reading method.

second important difference was that Jaffe, Shain, and Schuler (2018) split the corpus in two parts: in the first part, they evaluated which factors would contribute the most to the model and which reading zones were relevant. In the second part, they only tested the model with the parameters that were optimal in the first part. It prevented them from multiple testing and effecting a Bonferroni correction.

Jaffe, Shain, and Schuler (2018) found a highly significant result of mention count in their final model: when a mention refers to an entity that has been mentioned more often, reading times from the self-paced reading task are lower. However, they could not find convincing evidence for the recency-factors during their exploration of the first part of the corpus and therefore excluded it from the final model. Jaffe, Shain, and Schuler (2018) describe the effect they found as a small effect of mention count on reading time. They suggest that the study of natural reading data reflects better the influence that the factors have in every day language processing and might be exaggerated in the artificial settings of psycholinguistic experiments.

At the end of their article, Jaffe, Shain, and Schuler (2018), state nevertheless that the effect they found in the data could have to do more with surprisal than focussing. They state that if participants try to predict discourse entities, they could expect entities that were mentioned more often earlier in the text. In that sense, the mention count variable could be a measure of expectation rather than focussing.

This hypothesis is also supported by the fact that there is another factor — called story position, which represents how many percent of the story is read already — that is an extremely good predictor of reading time. This would mean that people read faster and faster when they progress in a story, a process that could also be explained by surprisal. But Jaffe, Shain, and Schuler (2018) show that when the story position is used as a baseline model, the mention count factor still improves the fit of the model over this baseline.

We think that this work is very interesting. First, it shows that it is possible to exploit natural reading data to answer questions about anaphora and coreference resolution. Second, the discussion about the role of surprisal and focussing effects is also very relevant for our thesis. Third, the Natural Story Corpus could also be an interesting resource for the purpose of our research questions, even though it should be kept in mind that it is not strictly speaking natural text corpus. And fourth, it discusses the difference in effect size between psycholinguistic studies using controlled items and reading corpus studies. However, we think that with respect to the cognitive hypothesis about coreference resolution, the same critique as the one we made about Dubey, Keller, and Sturt (2013) applies: we believe that the processing cost of coreference is determined by various factors. We therefore believe that the mention-frequency cost-metric is unlikely to explain all cost coming from coreference resolution.

4.5 Definition of our Cost Metric

In this section we present the cost metric we designed for pronoun resolution. To formulate a cost metric, we have to determine first what would cause difficulty in pronoun resolution. We hypothesize that the difficulty of finding the antecedent is determined by how much competition there is amongst antecedent candidates. Our hypothesis is the following: the higher the competition, the higher the processing cost to resolve the pronoun.

To capture this hypothesis, we propose a pronoun resolution cost metric based on the notion of entropy (please see Section 4.2.2 for more explanation). Entropy is a measure of ambiguity and we use it to measure competition. By using entropy as a cost metric, we hypothesize that the cognitive cost of pronoun resolution depends on the number of candidates and their degree of compatibility.

Our hypothesis about the cost of pronoun resolution is formulated in Equation 4.9. The cost of a pronoun, $C(pro)$, is calculated by using the probability distribution over the antecedent candidates A. The probability that every antecedent candidate a is the true antecedent of the pronoun is taken into account.

$$C(pro) = -\sum_{a \in A} P(pro = a) \cdot \log_2(P(pro = a)) \tag{4.9}$$

The idea behind our cost metric is thus fairly simple. But there are two big challenges. The first is to determine what an antecedent candidate is. The second is to obtain a probability distribution of coreference over the antecedent candidates.

We propose to use NLP-systems designed for pronoun resolution to both find the antecedent candidates and the probability distribution over them. We argue that these systems are designed to be robust and of large coverage. Therefore, our cost metric can implicitly take into account many features: all the probabilities come from the NLP-system that uses multiple features. We hope that because of this, our entropy cost metric is more plausible than cost metrics that only take one feature into account.

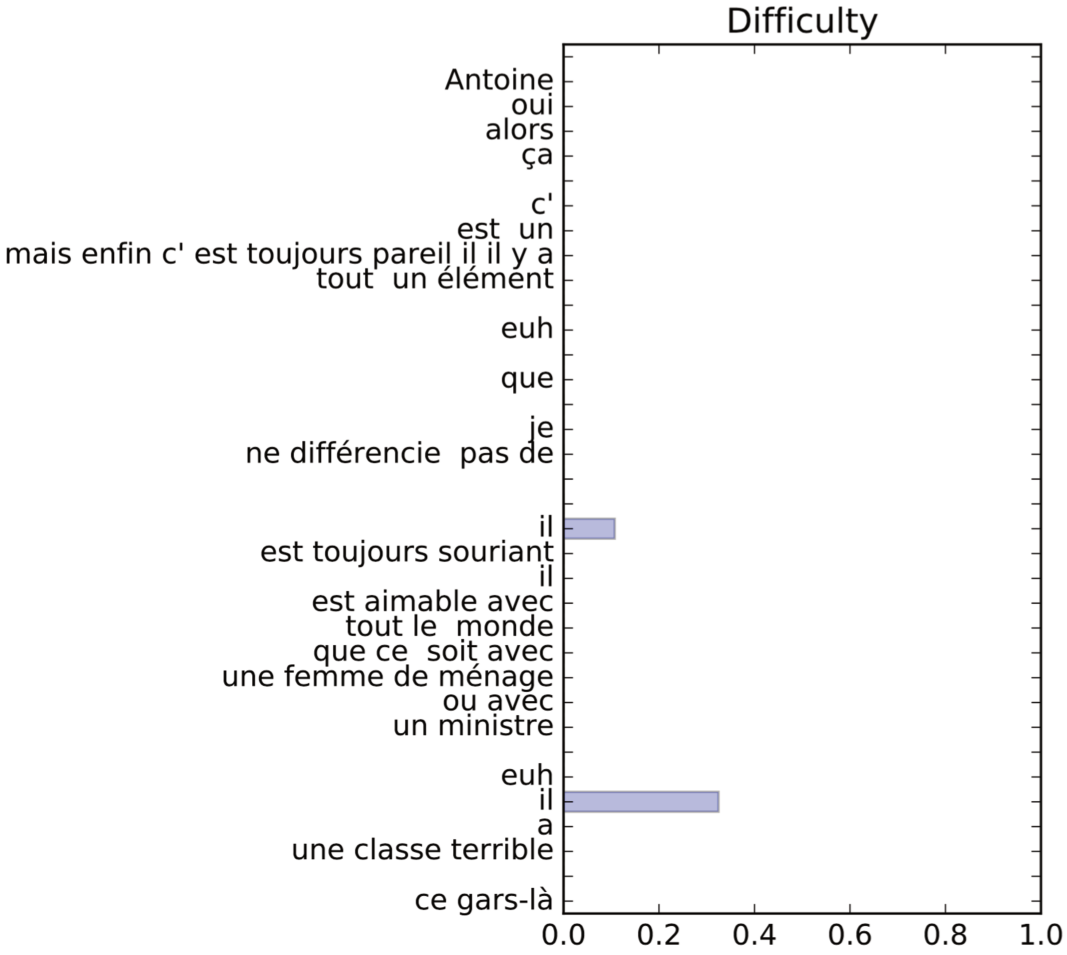
It is important to bear in mind that the maximal entropy $H_{max}(X)$ increases along with the number of possible outcomes of the random variable X. Therefore, it is likely that when there is a long discourse, pronouns further in the text systematically obtain a higher entropy score than those in the beginning of the text. Further away in the text, there are always more antecedent candidates. We therefore propose to normalise the entropy metric (see Section 4.2.4) or to use relative entropy (see Section 4.2.3) when it is tested on longer texts.

We illustrated the idea of our cost metric in Seminck (2016), a preliminary study on our cost metric. We developed a pronoun resolution system for French and tried to estimate normalised entropy for third person personal pronouns. We trained the resolution system and demonstrated the entropy metric on the largest French corpus annotated with coreference: the ANCOR-Corpus (Muzerelle et al., 2014) — an oral corpus of about half a million tokens.

Figure 4.2 illustrates how the prediction of the pronoun resolution cost metric looks on the corpus. For every personal third person pronoun, a score is emitted that reflects resolution difficulty. The idea is to investigate whether this prediction is a good model of processing cost coming from human pronoun resolution. However, the cost of pronouns predicted on the ANCOR-Corpus was very difficult to evaluate: there are no measurements of human processing cost available for it and because of its oral nature it would be difficult to collect. Therefore, the work presented in Seminck (2016) only has the purpose of illustrating the idea of the entropy cost metric.

FIGURE 4.2: Diagram in which the predictions of the normalised entropy costs metric of two personal third person pronouns in the ANCOR Corpus are illustrated. A translation of the transcription can be found in Gloss (7). Disfluencies and pauses are marked by blanks in the diagram and by ... in the transcription and the translation. The two pronouns 'il' (*he*) in this text fragment received a difficulty score based on our cost-metric. Note that the other pronouns, such as 'ça' and other 'il' are not scored because they are not personal pronouns or non-referring pronouns.

.



(7) a. Antoine, oui, ça... c'est un, mais enfin, c'est toujours pareil, il il y a tout un élément... euh... que... je ne différencie pas de... il est toujours souriant, il est aimable avec tout le monde, que ce soit avec une femme de ménage ou avec un ministre... euh, il a une classe terrible... ce gars-là

b. Antoine, yes, that... it's a... but yes, it is always the same, there there is a whole thing... er... that... I do not distinguish from... he is always smiling, he is nice to everybody, whether it is a cleaning lady or a minister... er, he is super classy... that guy

We will discuss more thoroughly the implementation details in the next two sections in which we present two experiments in which we tested the entropy cost metric. The first experiment compares the predictions of the entropy cost metric to the findings of Crawley, Stevenson, and Kleinman (1990) and Smyth (1994) presented in Chapter 2. The second experiment evaluated the entropy-metric on the Dundee Corpus.

4.6 Evaluation on Psycholinguistic Items

In the first experiment in which we tested the entropy-based cost metric we looked whether it was possible to simulate self-paced reading times recorded in the experiment of Crawley, Stevenson, and Kleinman (1990), discussed in Section 3.4.[5] We use our pronoun resolution system trained on the Ontonotes Corpus (Pradhan et al., 2011) (please see Section 3.3.4 for a full description of this system) to estimate a probability distribution over the antecedent candidates.

An important question is: how can we use the output of such a resolution system to obtain a probability distribution? Remember that the system was based on a logistic regression model that for two mentions m_i and m_j estimated a probability that these two are coreferent. We decided on a logistic regression classifier because we wanted to be able to interpret the features of the model. But what we are looking for now is a probability distribution over all entities that are candidates. Our resolution system does not provide a distribution over all candidates and works on the level of mentions, not on the level of entities. We therefore had to decide about a procedure to adapt the output from our resolution system to fit our needs.

To perform pronoun resolution on the experimental items, we first annotated all items with coreference information and syntactic trees. The coreference annotation was necessary to have a vision on the entities (coreference chains) in the items and not only the mentions. An example of the coreference annotation can be found hereunder, where Item (8) is followed by its coreference annotation.

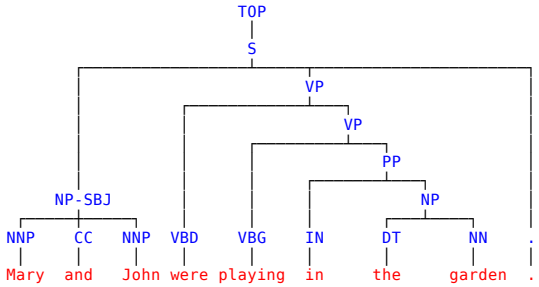
(8) Mary and John were playing in the garden. One of their classmates, Evelyn, tried to join in their game. Mary pushed John and Evelyn kicked him.

```
<DOC DOCNO="crawley_1">
<TEXT PARTNO="000">
<COREF ID="3" TYPE="IDENT"><COREF ID="subj" TYPE="IDENT" GENDER="f">Mary</
    COREF> and <COREF ID="obj" TYPE="IDENT" GENDER="m">John</COREF></COREF>
     were playing in <COREF ID="4" TYPE="IDENT" GENDER="n">the garden</
    COREF>. <COREF ID="5" TYPE="IDENT" GENDER="f"><COREF ID="6" TYPE="APPOS
    " GENDER="f">One of their classmates</COREF>, <COREF ID="6" TYPE="APPOS
    " GENDER="f">Evelyn</COREF></COREF>, tried to join in <COREF ID="7"
    TYPE="IDENT" GENDER="n"><COREF ID="3" TYPE="IDENT">their</COREF> game</
    COREF> . <COREF ID="subj" TYPE="IDENT" GENDER="f">Mary</COREF> pushed <
    COREF ID="obj" TYPE="IDENT" GENDER="m">John</COREF> and <COREF ID="6"
    TYPE="IDENT" GENDER="f">Evelyn</COREF> kicked <COREF ID="target" TYPE="
    IDENT" GENDER="m">him</COREF> .
</TEXT>
</DOC>
```

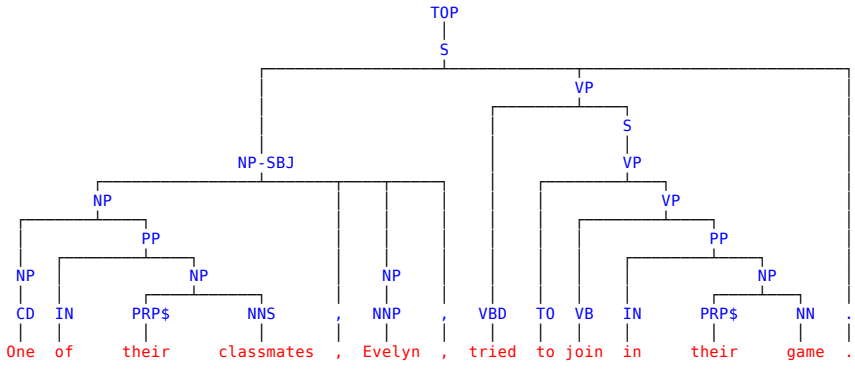
[5] Unfortunately, in Smyth (1994)'s experiment (also discussed in Section 3.4), no measure of processing cost was taken, so we could not evaluate our cost metric on their experimental items.

The syntactic annotation was necessary to provide the pronoun resolver with information about syntactic function features. We ran the Stanford Parser (Klein and Manning, 2003) on the items and corrected them manually. An example of the syntactic trees of the fragment above is:

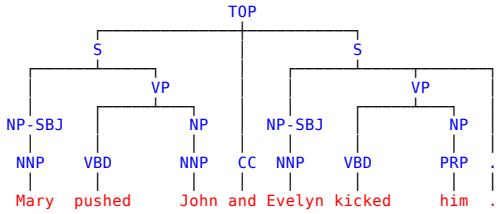
```
(TOP (S (NP-SBJ (NNP Mary)
                (CC and)
                (NNP John))
            (VP (VBD were)
          (VP (VBG playing)
                (PP (IN in)
                (NP (DT the)
                 (NN garden)))))
    (. .)))
```




```
(TOP (S (NP-SBJ (NP (NP (CD One))
        (PP (IN of)
          (NP (PRP$ their)
          (NNS classmates))))
      (, ,)
      (NP (NNP Evelyn))
      (, ,))
    (VP (VBD tried)
      (S (VP (TO to)
          (VP (VB join)
            (PP (IN in)
              (NP (PRP$ their)
               (NN game)))))))
    (. .)))
```




```
(TOP (S (NP-SBJ (NNP Mary))
    (VP (VBD pushed)
      (NP (NNP John))))
    (CC and)
        (S (NP-SBJ (NNP Evelyn))
    (VP (VBD kicked)
      (NP (PRP him)))
    (. .)))
```



To obtain a probability distribution over antecedent candidates represented as entities, we applied the following steps:

- From our pair-wise resolver (see Section 3.2.4 for an explanation), we got the coreference scores $P((m_i, m_j) = \textit{coreferent})$ between every preceding mention in the text and the pronoun that needed to be resolved.
- We then grouped the preceding mentions by their coreference chain. Because our resolution system did not build coreference chains, this information was taken from the manual annotation of the psycholinguistic items.[6]
- Each entity got the score of its highest scoring mention, this decision was inspired by the work of Luo et al. (2004).

[6]We made the strong assumption that recovering the coreference chains in the psycholinguistic items is rather easy and does not cause much processing cost.

- We considered all the entities that obtained a score >0.5 as antecedent candidates. We did not take into account negatively classified mentions as candidates. The reason was that we noticed that because of the architecture of our model, a logistic regression classifier working with a limited number of features, mentions that are classified negatively do not obtain a score close to zero. There are nearly always some features present in the mention pairs that augment the score, leading to scores of 0.3, or 0.4, for example. A score of 0.4 leads to an entropy increase as big as a score of 0.6: which is a positive result. We therefore concluded that negatively classified mentions did not add much to the competition there was between antecedent candidates and could therefore be excluded.

- We added a score for the 'empty' candidate (*i.e.* a probability for the event that the pronoun did not have an antecedent). We also followed Luo et al. (2004) in the assignment of probability to the empty candidate: it was given a probability equal to 1 minus the score of the highest scoring mention.

- To form a probability distribution over the antecedent candidates, we used the technique described in Luo et al. (2004): a probability distribution over the chains was formed by dividing the probability for each chain by the probability mass of all the chains in the distribution.

The entropy we used to evaluate our metric was calculated over this distribution. The procedure of transforming the output of the pronoun resolution system into a distribution over antecedent candidates represented as entities is illustrated in Table 4.3.

TABLE 4.3: We illustrate here how we obtained a probability distribution over antecedent candidates using a made up example. The pronoun *it* from the text in Example (9) has to be resolved and all preceding mentions in the text are reported under m_i. First $P(m_i)$ is output by the resolver and indicates the probability that m_i is coreferent with *it*. The empty candidate gets the score of 1 minus the highest scoring mention (hereabove: 1 - 0.95 = 0.05). Second, each mention is associated to its corefference chain c_i. Each chain gets the probability of its highest scoring mention, reported under P(c_i). Third, a probability distribution is forged from all candidates having a $P(c_i) > 0.5$ and the empty candidate. This is done by dividing the scores under $P(c_i)$ by the total probability mass of the maintained candidates (hereunder: 0.95 + 0.7 + 0.05). The result is a probability distribution, reported as *P(dist)*. Entropy is calculated on this distribution.

(9) The box was nice. It was great for the cat to sleep in, so it did not hesitate and jumped inside. Bob came home. He picked it$_{resolve}$ up.

m_i	P(m_i)	c_i	P(c_i)	P(dist)	Entropy
The box	0.95	*{The box, It}*	0.95	0.56	1.15
It	0.85				
the cat	0.7	*{the cat, it}*	0.7	0.41	
it	0.6				
Bob	0.01	*{Bob, He}*	0.2	–	
He	0.2				
∅	0.05	*{∅}*	0.05	0.03	

4.6.1 Items

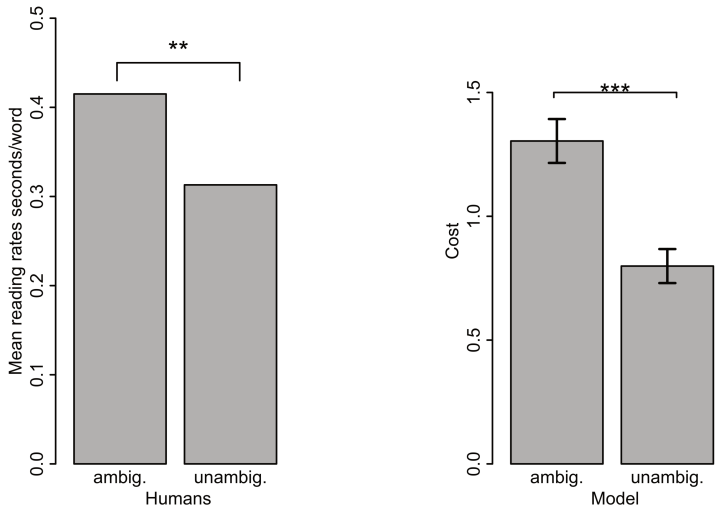
We ran our model on all experimental items of Crawley, Stevenson, and Kleinman (1990). Crawley, Stevenson, and Kleinman (1990) reported self-paced reading times of the last sentence of the experimental items. There were ambiguous sentences (these sentences were ambiguous because the pronoun had two antecedent candidates that were compatible in gender and number) and unambiguous items (see Section 3.4.2 for more details about the items). For all these items we calculated the entropy metric (every item contained one pronoun).

4.6.2 Results

We found that our cost metric was able to mirror certain reading times attested in the self-paced reading experiments. Crawley, Stevenson, and Kleinman (1990) reported a significant difference between the ambiguous and the unambiguous condition in an overall variance analysis of the data.[7] The model also shows this difference. When we performed an analysis of variance on a by-item basis, the factor of ambiguity was significant ($F = 299.5$, $df = 1, 39$, $p < .001$). In Figure 4.3 the predictions of the model and the actual experimental reading times are plotted against each other.

[7]We do not report the F-statistic here, because only the statistics for a by-subject analysis were reported.

FIGURE 4.3: The model's prediction of processing cost against the reading times per word recorded by Crawley, Stevenson, and Kleinman (1990) for the ambiguous and the unambiguous condition. For the cost predicted by the model 95% confidence intervals are given. The result reported for the humans is significant on, at least, an 0.01 level, but we had not engouh information to report error bars. Significance codes: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1.



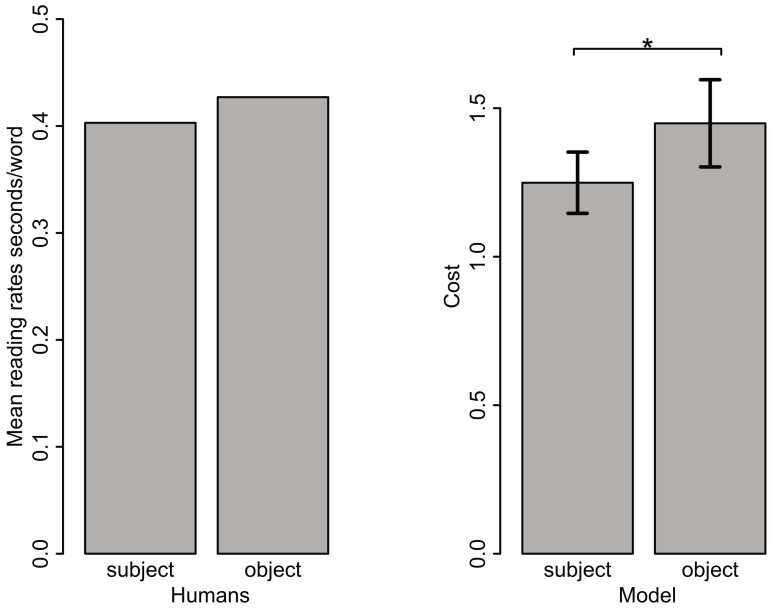
Crawley, Stevenson, and Kleinman (1990) also compared reading times between the subject and the object assignment inside the ambiguous and the unambiguous condition. They found faster reading times for subject assignment in the ambiguous condition, but this effect only showed in an analysis by participants and not by items ($F_1 = 8.53$, $df = 1{,}47$, $p < .01$; $F_2 < 1$). They did not find significant effects in the unambiguous condition, nor in the analysis by participants, nor in the analysis by items ($F_1 = 1.55$, $df = 1{,}47$, $p > 0.05$; $F_2 = 1.08$, $df = 1{,}39$, $p > 0.05$). Like Crawley, Stevenson, and Kleinman (1990), our model also showed a significant difference between subject and object assignment in the ambiguous condition ($F = 4.23$, $df = 1, 38$, $p < .05$), but in an by-item analysis[8]. For the unambiguous condition however, our results do not match Crawley, Stevenson, and Kleinman (1990)'s: we found a significant effect for the by-item analysis[9] ($F = 24.43$, $df = 1, 33$, $p < .001$). In Figure 4.4 the results for the subject and object assignment are plotted.

[8]The results of the model cannot be evaluated on a by-subject basis because there are no subjects.

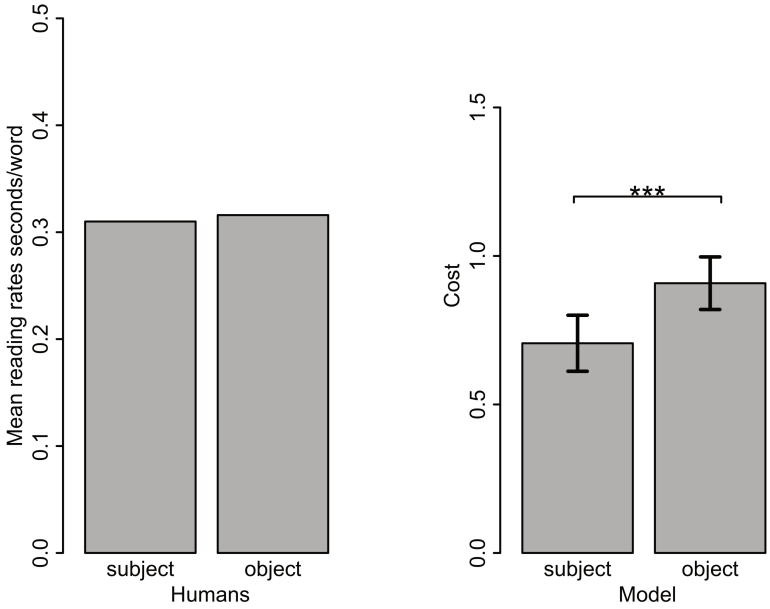
[9]In this analysis, items for which the resolver responded *None* were treated as missing values.

FIGURE 4.4: The model's prediction of processing cost against the reading times per word recorded by Crawley, Stevenson, and Kleinman (1990) for subject and object assignment in the ambiguous and the unambiguous condition. For the cost predicted by the model, 95% confidence intervals are given. We did not report them for Crawley, Stevenson, and Kleinman (1990) because we did not have access to the necessary information. The difference between the subject and object choices in the ambiguous condition (4.4a) was reported significant in a by-subject analysis, but not in a by-item analysis. For the unambigous condition, Crawley, Stevenson, and Kleinman (1990) did not find siginificant differences in reading time between the subject and the object choices. Significance codes: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1.

(A) Ambiguous Condition



(B) Unambiguous Condition



4.6.3 Discussion

Our experiment showed an example of the implementation of the entropy cost metric and an evaluation of it. The results are mixed. On the one hand, it seems that our model is capable of mirroring the reading times of ambiguous versus unambiguous items and the reading times of items with subject and object antecedents in the ambiguous condition. On the other hand, in the unambiguous condition we found an effect that was not observed in the human data. We thought of various reasons that can explain why we found these results.

The difference between the ambiguous and the unambiguous condition was simulated correctly by the model. However, it is probably a consequence of excluding mentions that received a probability score less than 0.5 as antecedent candidates. This rule could have included more 'events' into the distribution of the pronouns of ambiguous items, leading to a higher entropy. This does not invalidate our result but it is important to notice that it is possible that the result is more dependent on the decision what is a candidate rather than a direct test of our hypothesis that a higher competition leads to higher processing cost.

When we look at the difference between subject and object resolution in the ambiguous items, we can say that the weights that the model learned for the subject preference and the parallel function preference (the *syntactic path match* feature of the model) are quite accurate and that this explains why the model effectively displays the difference. This can therefore be interpreted as a genuinely positive result. Moreover, the rather big error bars in Figure 4.4a, shows that other features learned by the resolver also have an influence on the entropy cost metric.

The difference between the subject and object resolution in the unambiguous condition, however, is not modelled correctly. We could think of some explanations of the failure of the model here. First, we think that it can be explained by the strength of the gender and number features in our system. As the automatic gender and number feature assignment introduced some noise in our data, we think our model estimated these features lower than they should be.

A second reason, also compatible with the first reason, is that the way that we forged a probability distribution is not suited. By excluding entities with a score lower than 0.5, we might put too much probability weight on the remaining antecedent candidates and this might exaggerate their scores, amplifying entropy for object choices with respect to subject choices. If we left the low-scoring entities in the distribution, the effect would be weaker.

A third reason we can think of, is that there might be an effect between subject choices and object choices in unambiguous sentences, but this is not captured by the self-paced reading task formulated by Crawley, Stevenson, and Kleinman (1990). Indeed, only reading times of whole sentences were reported and it can be the case that the effect only shows very locally.

In conclusion, the results we obtained encourage us to continue to investigate the entropy cost metric: two out of three results have rather good models. However, we have to be more careful about the manner in which we establish a probability distribution. It would be better to have directly a full probability distribution on antecedents candidates, instead of needing to forge one. We therefore propose to use a ranking approach in the next experiment presented in Section 4.7. The treatment of negatively classified mentions also need rethinking: rather than excluding them directly, it would be desirable if they could get probabilities close to 0. In the

next section, we present an experiment in which we take these considerations into account.

4.7 Evaluation on the Dundee Corpus

In this section we present how we tested the entropy cost metric on the Dundee Corpus (Kennedy, Hill, and Pynte, 2003). We explain how we modelled the reading of the anaphoric pronouns in the APADEC resource (Seminck and Amsili, 2018); see Section 2.4.1 for more explanation about this resource. In Chapter 2 we already used the Dundee Corpus. We modelled first pass reading time for anaphoric pronouns making use of pronoun resolution biases described in the psycholinguistic literature.

During this experiment, we encountered various problems when building models of pronoun resolution with this data. One of the problems was data-sparsity: only ten participants read the Dundee corpus, leading to few observations per pronoun. A second problem was that we could not implement a satisfactory random factor structure. As we tested many factors of influence on pronoun resolution at once, there was not enough data to support the use of random slopes. Moreover, we had to deal with data-transformation and faced convergence problems for other reading times than the first pass reading time. Finally, we had to face a problem of multiple testing: we could not determine one single region of interest that would capture the influence of all factors and therefore defined multiple zones of interest. But the statistical effects we found were too weak to resist a correction for multiple testing.

In this section, we try to solve these problems. As a first step, we model the processing cost of pronoun resolution only with the entropy cost metric, instead of using various predictor factors of pronoun resolution as in the experiment presented in Section 2.5. The idea behind our cost metric is that the pronoun resolution system that provides the probability distribution does take into account these factors. The first advantage of this approach is that it enables us to use an adapted random feature structure in the mixed effects model. As a second advantage, the use of one pronoun resolution factor in the reading model leads to less reading zones. The multi-factor model from Chapter 2 showed effects from different factors in different reading zones. In the experiment presented in this section, we have only one factor. Therefore, the chances are high that we do not need as many zones of interest.

In Section 4.7.1, we explain how we faced the challenge of calculating a probability distribution over antecedent candidates. We present an NLP ranking system that we used to obtain a better probability distribution. Then, in Section 4.7.2, we present the implementation of our entropy cost metric to answer the questions of what is considered as a candidate and how the metric is adapted to the longer discourse present in the texts of the Dundee corpus. Then, in Section 4.7.3, we discuss what type of reading measurements we model. We propose another reading measurement that does help to increase the number of data-points per pronoun. In Section 4.7.4, we present the statistical models we used to investigate the cost metric. To overcome convergence problems, we choose a Bayesian inference model. We give a brief introduction to this type of statistics. Finally, we present the results (Section 4.7.5) and a discussion about this experiment (Section 4.7.6).

4.7.1 Resolution System

We used the current state of the art coreference resolver (Lee et al., 2017) (already referenced in Section 3.2.4) to provide us with a probability distribution over the antecedent candidates of the anaphoric pronouns in the Dundee Corpus. This system is characterized as an end-to-end coreference resolution system, which means that it does not need a preprocessing pipeline architecture that is commonly used for other systems. This makes the system very suitable for our purpose. Its strength lies in a process of mention-detection done simultaneously with the coreference resolution task. The system is intrinsically probabilistic, as it is based on a neural network ranker that outputs a probability distribution over different antecedents of a pronoun. Hence, the system can be characterized as a mention-pair resolution system. The final clustering algorithm to recover coreference chains is fairly simple: mentions are linked to the highest ranked mention — that can also be the empty candidate in case of non-coreferent mentions — and coreference chains are obtained by transitivity relations: when antecedent A is the highest scoring mention for pronoun B and also for pronoun C, a cluster of $\{A, B, C\}$ is formed.

4.7.2 Implementation of the Entropy Cost Metric

The outcome of Lee et al. (2017)'s system is a probability distribution over mentions that precede the mention that is resolved. However, we want to use a probability distribution over antecedent candidates, that is to say entity clusters. A first step is thus to regroup the mentions by their coreference chain. We simply take the coreference chains that are produced by the system as our antecedent candidates. The probability distribution over these candidates is obtained by summing the probability scores of all mentions inside the cluster (coreference chain).

We used relative entropy (Section 4.2.3) as a cost metric.[10] The documents in the Dundee corpus can be quite long. By using the relative entropy, the entropy metric is not always higher at the end of the document than at the beginning because of the number of antecedent candidates.

4.7.3 Reading Metric

We propose to look at fixations on the pronoun — or the absence of them. As discussed in Section 2.5.3, we only have 4 data-points on average for every pronoun if we look at reading times. The Dundee Corpus was only read by ten participants and in about 60% of the cases, candidates skip pronouns. But when we look at fixation as a boolean variable, we have ten data-points for every pronoun: they either fixated it or not. This new reading metric helps to overcome data-sparsity. As a second advantage, this measure also prevents to have multiple zones of interest.

4.7.4 Statistical Model

For this experiment, we used generalized linear models with random effects. Our implementation uses the Bayesian inference framework in R. The reason for shifting

[10] We decided to use relative entropy and not normalised entropy because we found better descriptions of relative entropy than of normalised entropy in the literature (Thomas and Cover, 2006). Besides, both metrics being similar, the normalised entropy is expected to behave in a way that is similar to the relative entropy.

to a Bayesian framework was motivated by convergence problems we had with the frequentist framework that was due to a failure in the parameter estimation of random variables. Bayensian models display in general fewer estimation problems (Eager and Roy, 2017) and are better adapted to small data (Muthén and Asparouhov, 2012). They also do not need correction for multiple testing (Gelman, Hill, and Yajima, 2012). In Bayesian statistics, the concept of false positive does not play a role. Instead of evaluating the null-hypothesis, as in the frequentist framework, the likelihood of the data is estimated directly. In other words: in a frequentist framework, the α-level of testing is the chance the researcher takes to mistakenly reject the null-hypothesis. It is problematic when α is increased because it means that there is a higher chance that the data was produced by random errors instead of a real effect. However, in the Bayesian framework, the null hypothesis does not play a role and therefore the chance to mistakenly reject it is irrelevant. In the following paragraph, we briefly explain how Bayesian statistics work following the explanation of Downey (2013).

Bayesian Models

The core of Bayesian statistics is Bayes' Theorem (Bayes and Price, 1763). Downey (2013) explains this theorem the following way. Imagine that we have two coins, A and B and we want to know the probability that if we flip them we get two heads. If A and B both have a probability of 0.5 to give heads, we know that the probability of obtaining two times heads is $0.5 \cdot 0.5 = 0.25$. Note that A and B are independent from each other: if we get heads for A it says nothing about what would happen to B. Therefore, for two independent random variables, we can say that:

$$P(A \wedge B) = P(A) \cdot P(B) \tag{4.10}$$

But many variables we can be interested in are not independent. For example, if we are interested in the two events: *it will rain today* (C) and *it will rain tomorrow* (D), the two events are not independent because rain can typically last some days. Therefore, if C is true, there is more chance that D is true. We can say that $P(C \wedge D) > P(C) \cdot P(D)$ in this case. For dependent variables, it holds that:

$$P(C \wedge D) = P(C) \cdot P(D|C) \tag{4.11}$$

$P(D|C)$ is the conditional probability of D on C. It gives the probability that D happens, given that C is true. Crucially, in $P(C \wedge D)$, C and D are interchangeable: *it will rain today* and *it will rain tomorrow* is the same thing as *it will rain tomorrow* and *it will rain today*, therefore:

$$P(C \wedge D) = P(D \wedge C) \tag{4.12}$$

This leads automatically to the following formula, which forms the basis of Bayes' Theorem.

$$P(C) \cdot P(D|C) = P(D) \cdot P(C|D) \tag{4.13}$$

Bayes' Theorem can then be derived easily:

$$P(D|C) = \frac{P(D) \cdot P(C|D)}{P(C)} \tag{4.14}$$

If we now switch to the case of hypothesis testing, we can say that for a given hypothesis, H, we want to calculate the probability that it is true, given some evidence E. This can easily be rewritten as Bayes' Theorem:

$$P(H|E) = \frac{P(H) \cdot P(E|H)}{P(E)} \tag{4.15}$$

$P(H)$ is what is called the prior distribution. It can be referred to as the beliefs we have about an hypothesis, before any evidence is taken into account. $P(H|E)$ is called the posterior and this is the value that we want to know. $P(E|H)$ is the likelihood that the evidence comes out, if the hypothesis H were correct. Finally, $P(E)$ is called the normalising constant. It is the probability of the data (*evidence*) under any circumstances.

When we do Bayesian statistics, we start with a hypothesis, which is represented by the prior. Or, more generally, we start with a collection of hypotheses to which we attribute probability mass, this is referred to as the prior distribution (Downey, 2013). We illustrate this with an example given by Downey (2013).

Imagine a game that features a 4-sided die, a 6-sided die, a 8-sided die, a 12-sided die and a 20-sided die. Each die has numbers on it going from 1 up to the number of sides it has. If somebody took one of the dice and was rolling it, not saying to us which die they took but only the outcomes of the roles, we could guess the die by using Bayesian statistics.

In the beginning, before any rolls, we attribute uniform probability to every die: that is to say $\frac{1}{5}$. This is our prior distribution, which will be updated according to the evidence (the outcome of the die rolls) that will be presented.

Imagine that the first roll gives the number 6. We use this information to update the distribution. For every die, the probability of that die obtaining the value of 6 is multiplied with the probability of the prior distribution. After this step, these numbers are normalised by the remaining probability mass. After the roll of 6, the 4-sided die is immediately excluded, because it has a 0.0 probability for the value 6. The evolution of the distribution after rolling two more numbers — 4 and 8 — can be found in Table 4.4.

TABLE 4.4: An example of how a prior distribution can be updated using data from Downey (2013). There are five dice, one 4-sided, one 6-sided, one 8-sided, one 12-side and one 20-sided with numbers ranging from 1 till the number of sides. Somebody took one of the dice — without telling us which one. We have to guess which die they took only by obtaining the numbers that were rolled. There were 3 rolls and it got us the numbers [6,4,8]. Using Bayesian statistics, after each roll, the probabilities that each die was chosen are updated.

	number rolled			
n-sides of die	-	6	4	8
4	0.20	0.0	0.0	0.0
6	0.20	0.392	0.526	0.0
8	0.20	0.294	0.296	0.735
12	0.20	0.196	0.131	0.218
20	0.20	0.118	0.047	0.047

In the example above, we see that the prior distribution is of influence on the posterior distribution, especially in the beginning. In the example we choose to use a uniform, or uninformative, prior, because we have no reason to believe that one die is more likely to be picked than another. However, if for example our problem was to determine the probability that a coin falls on heads, a uniform prior is less likely. It is very unlikely that it falls 0%, or 100% of the times on heads. It is more likely that it falls 50% of the times on heads. Therefore, another distribution is suited for this problem that attributes more probability mass to values between 0.40 and 0.6 and less to the edges. It is nevertheless necessary to point out that if there is enough data, the posterior distribution will converge, regardless of the prior (Downey, 2013). However, it is assumed that it is a better idea to give some shape to the distribution if there is a hypothesis about its shape.

When we work on reading time problems, the Bayesian statistics is similar to the example of the dice. When we work with a regression model, we suppose prior distributions on the values of the parameters θ of the model. These distributions are updated with data until we obtain a posterior distribution that represents the values of θ, when all the data has been taken into account.

To decide whether a factor in the model is of influence, we look at the credible intervals. A credible interval is an interval of values in which we find a given proportion of the data, for example 95%. It means that there is a 95% chance that the value of θ lies between the two values of the interval. Note that often a 95% interval is chosen to look a bit comparable to the 5% α-level in frequentist statistics. It should be kept in mind, however, that it is not exactly comparable. The α-level expresses the chance that the null-hypothesis has been falsely rejected, whereas the 95% credible interval in Bayesian statistics corresponds to a 95% chance that the value of the parameter falls into the interval. It is therefore accepted in Bayesian statistics to also look at lower credible intervals: when you look at a 90% credible interval it does not mean that there is a 10% chance that the outcome is based on a randomly obtained result, as it would be the case for a 10% α-level in frequentist statistics.

Implementation

To implement the Bayesian models, we used the brms (*Bayesian Regression Models using Stan*) package (Bürkner, 2017) in R (R Development Core Team, 2008). We called the function brm. We set the family argument to a bernoulli distribution, because we tested whether the pronoun was read or not. We tested with weakly informative priors. We used a weakly informative normal distribution for the coefficients' and the intercepts' priors and a *lkj* prior — a weakly informative prior for correlation matrices (Lewandowski, Kurowicka, and Joe, 2009) — for the correlation parameters. We introduced random intercepts for participants and items (pronoun instances of the Dundee Corpus) on the entropy metric. We also used random slopes for participants on the entropy metric. We did not include random slopes for the items on the entropy metric, because every item has only ten data-points. Besides, the number of items is very high and it is thus likely that the model does not suffer much from 'idiosyncratic behaviour' of individual items.

We introduced the following control factors into our model:

- Length of characters of the zone of interest;
- Log frequency of the word in the zone from the British National Corpus;
- Whether there was a comma inside the zone;
- Whether there was a 'hard' punctuation inside the zone (period, exclamation mark, or question mark);

Hereunder is the code of our model. The arguments *control = list(adapt_delta) = 0.99* and the number of iterations, *iter*, ensure the quality of the sampling.

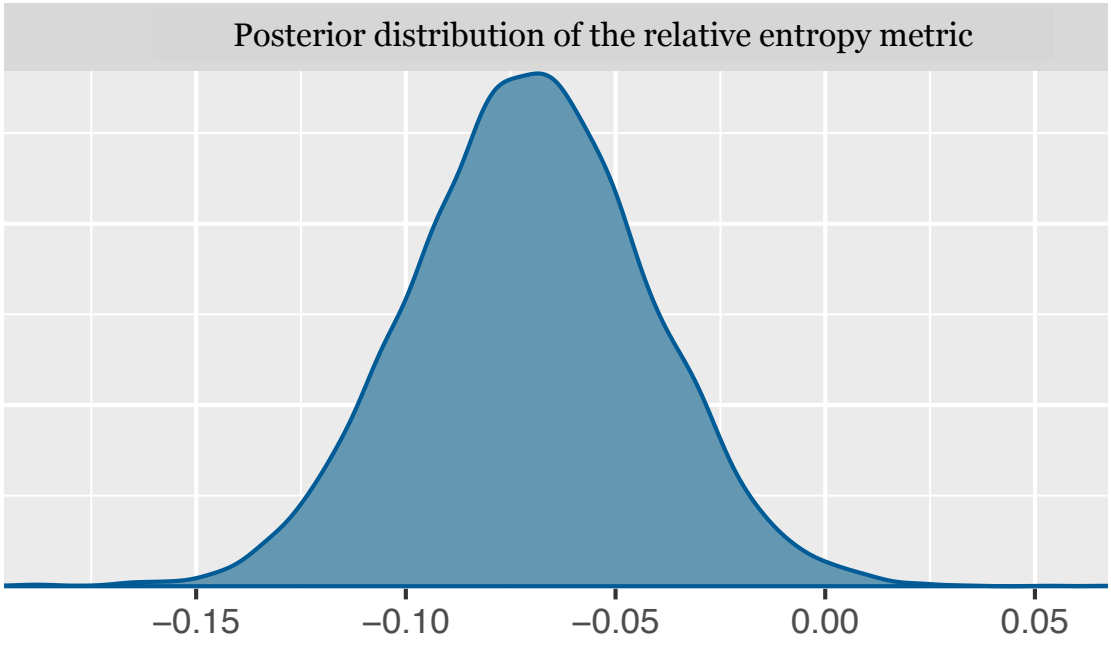
```
prior <- c(set_prior("normal(0,1)", class="b"), set_prior("normal(0,1)",
    class="sd"), set_prior("lkj(2)", class="cor"))

mod_brm_rel_ent <- brm(read ~ length_in_chars  +  frequency_bnc + comma  +
     hard_punct + rel_ent +  (1 + rel_ent|participant)  +  (1 |
    dundee_tokens) , family = bernoulli, data=pro, prior=prior, chains = 6,
     iter = 3000, control = list(adapt_delta = 0.99))
```

4.7.5 Results

In this section, we present the output of the model testing the relative entropy cost metric. The 95% credible intervals showed that, as expected, a lower BNC-frequency, a longer length in characters, and attached comma lead to a higher percentage of fixations on the pronoun. The presence of a period, an exclamation mark or a question mark (the *hard punct* factor) did not show a clear sign. The posterior distribution of the metric of relative entropy (illustrated in Figure 4.5) has a credible interval that does not contain zero: $[-0.13, -0.01]$. It is placed on the negative side, meaning that when there is a smaller distance between the maximal entropy distribution an the actual probability distribution output by the resolver, there is a higher chance of fixating the pronoun. This result is thus in the direction that we predicted: when there is more competition amongst the antecedent candidates, there is a higher chance that the participants read the pronoun. The full model as output by R can be found in Appendix A.

FIGURE 4.5: The posterior distribution of the relative entropy metric.



4.7.6 Discussion

We observed that the relative entropy cost metric has an influence on whether the pronoun is read of skipped. It would be interesting to see whether the entropy cost metric could be validated on other types of psychological measurements. For this experiment, we took the very simple information whether the pronoun was fixated or not. Our choice was motivated by the difficulty of determining a zone of reading time on the one hand and on the other hand by the fact that this reading information leads to more data than for example classical reading times, such as the first pass reading time.

4.8 Conclusion

Information Theory provides us with tools to measure information. In research on human syntactic processing, the quantity of information conveyed by syntactic structures affected the cognitive load: structures conveying more information demand more processing cost. The literature shows that this type of predictions does not only hold for syntactic structures but can be generalized to other domains, such as semantics or discourse.

We believe that information must also play a role in pronoun resolution. We hypothesized that competition is important for pronoun resolution: the more competition there is amongst the antecedent candidates of a pronoun, the more processing cost it will demand. We proposed to measure the competition with the information cost metric of entropy, which is often described as a measure of ambiguity. Entropy is measured over a probability distribution and it is maximal when all probabilities are equal. To measure the entropy of a pronoun, we thus estimate a probability

distribution over its antecedent candidates. We used NLP-systems of coreference resolution to obtain probability distributions over antecedent candidates.

We evaluated our hypothesis in two experiments. In the first experiment, we looked whether the entropy cost metric was a good model to simulate reading times of a self-paced reading experiment. In this experiment, the difference between ambiguous and non-ambiguous pronouns was studied together with the influence of the syntactic position of the antecedent. We used a simple resolution system to estimate the probability distribution over the antecedent candidates in the experimental items. The entropy cost metric was able to simulate the difference in reading times between ambiguous pronouns and unambiguous pronouns. It also simulated the lower reading times of ambiguous sentences when participants choose an antecedent in the subject position. But it failed to simulate the equal reading times of sentences featuring an antecedent in the subject or the object position in the non-ambiguous condition. We found the results encouraging but had reservations with respect to the manner in which the probabilities of antecedent candidates were calculated. Indeed, to be more certain about the predictions of the model, it was necessary to obtain a more reliable probability distribution.

In the second experiment, we evaluated the competition hypothesis on the Dundee Corpus. Because entropy is affected by the number of possible outcomes — in our case the number of antecedent candidates — we used the metric of relative entropy, that measures a 'distance' between the entropy and the maximal entropy. In that way, we prevented the pronouns at the end of a text from obtaining systematically higher scores. We found that the relative entropy is a factor of influence on human pronoun reading behaviour. We therefore conclude that competition amongst antecedent candidates of a pronoun is of influence on human pronoun resolution: more competition leads to more cognitive load.

Our experiments do not prove that competition is the only factor of influence on pronoun resolution. It just proves that models of human pronoun resolution must take this factor into account. It is plausible that prediction also has an influence on pronoun resolution. Expectations about which referent is mentioned next can have an influence on the resolution process: when the pronoun is encountered and resolved, it could be that resolution is speed up when the antecedent was the expected referent. The prediction aspect will be discussed further in Chapter 5.

Chapter 5

A Bayesian Model of Pronoun Resolution Evaluated on Corpus

5.1 Introduction

In this chapter we want to make a first step in testing existing theories of pronoun resolution on natural corpus data. We believe that testing on natural data is important for the development of theories because it verifies whether a theory is robust. The claims of theories are often broad: many theories aim at explaining language processing in general and, of course, this is not a problem. But the evidence for the theories often comes from psycholinguistic experiments in which the theory's influence is shown on a small number of phenomena. The phenomena are supposed to be generalizable over a larger pool of linguistic phenomena. But if the predictions of a theory also bore out to be of importance on uncontrolled data, it would be a much stronger proof of its robustness.

In the experiment that is presented in this chapter, we work towards a corpus account for a recent theory of pronoun resolution described in Kehler and Rohde (2013) and Kehler and Rohde (2018). This theory states that pronoun interpretation is influenced by *saliency* on the one hand, and by *world knowledge* on the other hand. Both these components have played an important role in theories about pronoun resolution. Kehler and Rohde's theory provides a probabilistic formula — based on Bayes' Theorem (Bayes and Price, 1763) — to calculate what specifies the relation between saliency factors and world knowledge factors.

Our choice to evaluate this theory was motivated by the fact that it features a quantitative model — which we could actually implement —, the fact that the theory takes into account multiple factors of influence and third, that until now all the evidence for it comes from psycholinguistic studies featuring hand-made items.

The challenge for us is to calculate the probabilities in the formula with respect to natural data. Indeed, whereas the experiments are designed in a way that it is obvious how to obtain a probability space, it is less evident on corpus data. The reason for this is that markers of saliency and world knowledge are introduced by hand in the experimental items used by Kehler and Rohde (2013). Obviously, this is not the case for natural data. Because saliency and world knowledge are rather complex and abstract concepts, measuring them on corpus data is a challenge.

To test Kehler and Rohde's theory on corpus data, we started with a very simple approach to estimate the probabilities. We annotated the coreference chains in an eye-tracking corpus, the Provo Corpus (Luke and Christianson, 2016), that also contained cloze task predictions: humans had to predict for each word in the text

what would be the next word. We used the cloze task data to estimate the probabilities needed to test the Bayesian model. It turned out that this simplistic method already yielded some interesting results: on the one hand we find some evidence for Kehler and Rohde's theory and on the other hand the data allows us to exploit the role of prediction in pronoun resolution. We therefore think that a more sophisticated method that is more costly in time and money will be worth the exploration in future work.

This chapter is structured in the following manner: we first discuss in depth Kehler and Rohde's theory. After that, we will discuss the implementation we did to investigate the theory on the data of the Provo Corpus (Luke and Christianson, 2016). Then, we will discuss the results we obtained. It must be kept in mind that they are preliminary results and that more precise methods should be used to estimate the probabilities with more accuracy. We finish the chapter with a discussion about the perspectives for further research.

5.2 A Bayesian Model of Pronoun Interpretation: K&R-Theory

In this section, we explain Kehler and Rohde's theory based on their article from 2013. From now on, we will refer to the theory using the initials of the authors: K&R-Theory.

In K&R-Theory, it is proposed that pronoun interpretation is driven by two type of processes: processes that have to do with information structure and processes that have to do with world knowledge. In the theory, the first type of processes is referred to as *centering-driven*, whereas the second type has the title of *coherence-driven*. The title of the first article describing the pronoun interpretation theory of Kehler and Rohde (2013) captures this idea quite well: *A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation*. However, in a more recent work about K&R-Theory, the theory is referred to as *A Bayesian Theory of Pronoun Interpretation*, a title that is less informative, but much shorter and therefore easier to handle. What is important here is on the one hand that both *Bayesian Theory* and *Reconcilliation between Centering and Coherence* refer to the same theory, and on the other hand, that the word *Bayesian* does not have much to do with the Bayesian statistics presented in Chapter 4, except that it also makes use of Bayes' Theorem (Bayes and Price, 1763).

The reason for K&R-Theory to suppose that two types of processes are at work in pronoun resolution is because some pronoun resolution behaviour can only be explained using both. An example is (1). For each of the sentences in this example, the referent that most native speakers of English would choose for the pronouns is given.

(1) a. Mitt narrowly defeated Rick, and the press promptly followed him to the next primary state. [him = Mitt]
b. Rick was narrowly defeated by Mitt, and the press promptly followed him to the next primary state. [him = Rick]
c. Mitt narrowly defeated Rick, and Newt absolutely trounced him. [him = Rick]
d. Mitt narrowly defeated Rick, and he quickly demanded a recount. [he = Rick]

Both examples (1-a) and (1-b) can be explained by the preference to resolve an ambiguous pronoun to the subject (see Section 2.2.1 for a more detailed description)

and example (1-c) can be attributed to the preference to resolve a pronoun to an antecedent that occupies a parallel syntactic position (see Section 2.2.2 for more information about the parallel function). However, for example (1-d), neither the subject bias, nor the parallel function can explain the preference for *Rick*. The reason for native speakers to choose *Rick* is world knowledge: the politician that looses a vote is more likely to ask for a recount.

According to K&R-Theory, the difference between Example (1-a) and (1-b) in which the subject bias plays a role on the one hand and example (1-c) in which the parallel function seems to be operational on the other hand, can be explained by Centering Theory (Grosz, Weinstein, and Joshi, 1995) — see Section 3.2.1 for a more detailed description of this theory.

On the contrary, example (1-d) cannot be explained by Centering, but it can be explained by a theory proposed by Hobbs (1979). This theory is based on coherence relations. Coherence relations are relations between sentences[1] that state how the two parts of text connect with each other in a discourse. Some examples are: *elaboration* and *causality*. According to Coherence Theory, an important part of text comprehension is inferring these coherence relations. This is done by applying world knowledge, such as in example (1-d). According to theories of pronoun resolution based on coherence relations, a pronoun is a free variable that becomes bound when the coherence relations of a text are established. Note that Coherence Theory cannot explain why the preferred interpretations of examples (1-a) and (1-b) are different: the semantic content is the same.

K&R-Theory is supported by completion studies from the psycholinguistic literature and the authors themselves. First, results in support of Coherence Theory are discussed. Studies with completion task stimuli like (2) are reported that serve to investigate whether participants refer more to the *source* semantic role of the sentences, which is *John*, or the *goal* semantic role: *Bill*.

(2) a. John passed the comic to Bill. He . . .
b. John was passing the comic to Bill. He . . .

The manipulation between stimuli (2-a) and (2-b) is the grammatical aspect between the perfective and the continuous aspect. In general, there is a preference for choosing the semantic role of 'source' (John), but there are important differences between stimuli (2-a) and (2-b). Because the perfect often has the interpretation of a completed action, there are more references to the semantic role of 'goal' (Bill) in (2-a) (43%) than in (2-b) (20%), where the action is still going on.

This finding suggests that it does not make much sense to talk about a preference for interpreting pronouns as referring to the source that can function as a heuristic, because the manipulation of the aspect is of such an importance.

An important observation made on this data is that the bias towards the source was very much dependent on the coherence relation that the participants employed to complete the stimuli. Moreover, not all types of coherence relations were employed in equal quantities: participants often used the *occasion* relation and the *elaboration* relation and not so often the *result* relation. Crucially, the bias to the source

[1] Actually, coherence relations are formulated between *discourse units*. It depends on the theory what is defined as a discourse unit. It can be defined as a sentence, a clause or a proposition. We keep it to sentences here, because this type of discourse unit is used in the experiments of Kehler and Rohde (2013).

was dependent on the employed relation. Therefore, K&R-Theory states that coherence relations are a driving force in pronoun interpretation.

A second important observation, that was already made in the literature by Stevenson, Crawley, and Kleinman (1994), is that participants do not choose the same proportion of pronouns to refer to entities in the subject and the object position. When stimuli without pronoun prompts, such as (3), were proposed, participants could decide on the referential forms. They used pronouns more often to refer to the syntactic subject and used more often a name to refer to other syntactic roles. This effect cannot be explained by coherence relations.

(3) John passed the comic to Bill. ...

According to K&R, the difference in results between example (2-a) and (3) are caused by the fact that they do not measure the same thing. They argue that stimuli like (3) measure the probability of mentioning a referent $P(\mathit{referent})$, whereas in (2-a) it is measured whether a referent is mentioned given that a pronoun has been used: $P(\mathit{referent}|\mathit{pronoun})$.

$P(\mathit{referent}|\mathit{pronoun})$ is the pronoun comprehension problem. Because it is a conditional probability, it can be decomposed by Bayes' Theorem (Bayes and Price, 1763); see Equation 5.1. On the right hand side of the decomposition of this conditional theory, we find $P(\mathit{pronoun}|\mathit{referent})$, which can then be seen as the probability of pronoun production. Therefore, K&R-Theory states pronoun interpretation is not merely the inverse of pronoun production.

$$P(\mathit{referent}|\mathit{pronoun}) = \frac{P(\mathit{pronoun}|\mathit{referent})P(\mathit{referent})}{\sum_{r \in \mathit{referents}} P(\mathit{pronoun}|r)P(r)} \tag{5.1}$$

The Bayesian formula states that the pronoun interpretation is influenced by the likelihood that a referent is mentioned next, indicated by $P(\mathit{referent})$, and by the conditional likelihood that a given referent will be pronominalized: $P(\mathit{pronoun}|\mathit{referent})$. This last probability can be seen as the production probability.

K&R-Theory states that $P(\mathit{referent})$ is determined by coherence driven factors and $P(\mathit{pronoun}|\mathit{referent})$ by information structure (e.g. topic and focus) of the discourse: the centering-driven factors. Various completion studies are reported to support these claims. We enlarge here on one of them to illustrate the evidence in favour of K&R-Theory.

In the study we picked, the claims about $P(\mathit{referent})$ and $P(\mathit{pronoun}|\mathit{referent})$ are investigated. To do so, items that feature implicit causality verbs and two first names of the same gender were created. Two manipulations on these items are proposed: a voice manipulation (active and passive voice) and a pronoun manipulation that consisted of giving a pronoun prompt or not after the sentence with the implicit causality verbs: participants are either forced or not to use a pronoun to complete the passage. An example of an item with all the manipulations is given in (4).

(4) a. Amanda amazed Brittany. She ...
b. Brittany was amazed by Amanda. She ...
c. Amanda amazed Brittany. ...
d. Brittany was amazed by Amanda. ...

K&R-Theory makes the following claims about these items:

1. When a pronoun prompt is used and the voice is passive (Item (4-b)), fewer participants will interpret 'She' as Amanda than when the voice is active (Item (4-a)). In pronoun production, there is a bias to refer to the subject with a pronoun, which is according to the Bayesian formula of influence on pronoun interpretation. When the implicated referent (Amanda) does not correspond to the preferred syntactic role (the subject), there will be fewer references to the implicated referent.

2. In general, more references are made to the subject than to the object when a pronoun is used. Therefore it is predicted that Item (4-a) will lead to more subject references than Item (4-c) and Item (4-b) more than Item (4-d).

3. The next mention bias is influenced by the coherence relation employed by the participants in their continuation. Because the intended referent is not in the subject position while a pronoun is used in Item (4-b), the attention is drawn away from it. This leads to fewer explanation relations (the relation that is evoked by the implicit causality verb) than in Item (4-a).

4. For the items without the pronoun prompt, Item (4-c) and Item (4-d), the pronominalization rate is affected by the saliency of the referent. In English, subjects in passive sentences are more salient than subject in active sentences. Therefore, it is expected to find a higher pronominalization rate for Item (4-d) than Item (4-c).

All these claims were confirmed by the experimental data. Moreover, $P(pronoun|referent)$ and $P(referent)$ were calculated using the no-pronoun prompts (Items (4-c) and (4-d)) and estimate subsequently $P(referent|pronoun)$ using Bayes' Theorem. The results could then be compared to the responses obtained from participants for the items with a pronoun prompt (Items (4-a) and (4-b)). The predicted biases of pronoun interpretation matched the attested biases very closely, as illustrated in Table 5.1.

TABLE 5.1: The predictions that the Bayesian Model makes about pronoun interpretation versus the attested biases of pronoun interpretation from K&R's completion experiment. The predicted values match the attested values closely.

	Active	Passive
Predicted	0.81	0.59
Actual	0.74	0.60

In a second article (Kehler and Rohde, 2018), other completion experiments are reported in support of K&R-Theory. In addition to these completion studies, it was evaluated whether $P(referent|pronoun)$ obtained by the Bayesian formula matched human responses better than $P(pronoun|referent)$, or $P(referent)$ on their own. The result was positive, suggesting that the Bayesian model is more adequate than models that assume that pronoun interpretation is the mere inverse of pronoun resolution and models that assume that only the mention bias is of influence.

5.3 Implementation

In our experiment, we wanted to test whether this model is also relevant for pronouns in natural texts and test the robustness of its predictions. The first challenge is to get estimates for the components of the formula: $P(pronoun|referent)$ and $P(referent)$. There are many possible ways to approximate these terms. In the proposition of Kehler and Rohde (2013), $P(pronoun|referent)$ is driven by considerations speakers make about the information structure of their utterances and could be calculated by the principles of Centering Theory. $P(referent)$ is the bias of mentioning a referent, regardless the referential form. Whether $P(referent)$ is mentioned, depends on the coherence relations of the discourse.

To respect these hypotheses, we could use NLP-implementations of Centering Theory and semantic models that predict which referent of a discourse is the most likely to be referenced next. However, we could also take a step back from these theories and first evaluate whether the Bayesian formula yields results by estimating $P(pronoun|referent)$ and $P(referent)$ on human data. Indeed, completion experiments using natural data could provide these probabilities. Using human responses to estimate these two probability terms does not allow to test whether Centering and Coherence are of influence on them, but the Bayesian proposition could still be examined.

We thus present a pilot experiment that investigates the Bayesian Theory of Pronoun Resolution based on human responses on natural data. We used the Provo Corpus (Luke and Christianson, 2016) that is distributed with cloze task data and eye tracking data. In the perspective section — Section 5.6 — we discuss how on the one hand the estimation of probabilities using human data can be enhanced and how, on the other hand, the probabilities could be estimated using automatic methods that are implementations of Centering Theory and Coherence.

5.3.1 Metrics

We tested three probabilities on the Provo Corpus:

- The Baysian model, that is to say $P(referent|pronoun)$ obtained with Equation 5.1.
- $P(referent)$, as an approximation of the expectation of the antecedent.
- $P(exact\ form)$, the probability that the participants in the cloze task guessed the pronoun right.

5.3.2 The Provo Corpus

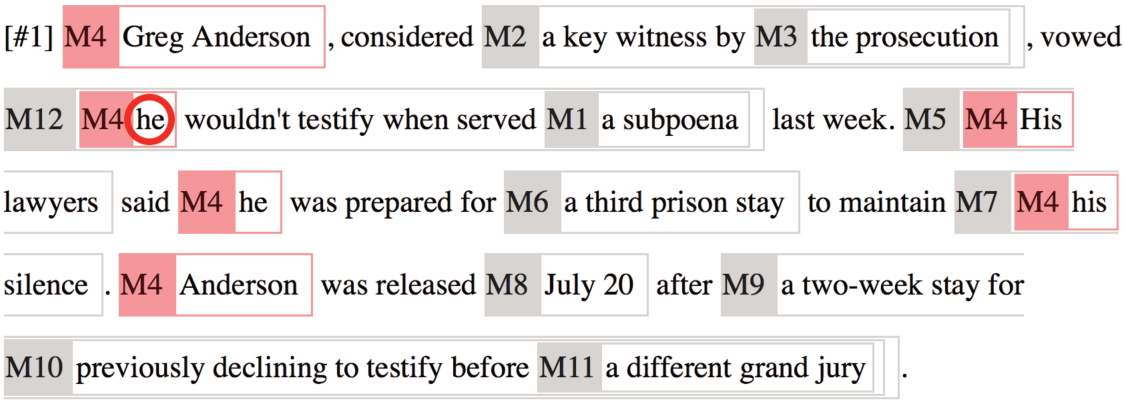
For this experiment, we are using the Provo Corpus (Luke and Christianson, 2016). This is a small corpus of natural text reading. It contains 55 short text passages and counts in total of 2 689 tokens. The particularity of this corpus is that it contains two types of human responses: eye-tracking data and cloze-task responses. The number of participants for both is quite high: 84 participants were eye-tracked and for each word in the corpus, except for each first word of the 55 passages, around 40 cloze task responses were collected.

The cloze task consist in making participants guess what the next word of a text will be. First, only the first word of the text is presented, and participants are asked what word would follow. The participants type in their answers, and the word that was in fact in the text appears, after which the participant is asked which word would follow next. This goes on until the end of the text fragment. As a first approach to the Bayesian model, we decided to use the cloze-task responses of the data to estimate the probabilities of reference and pronoun production. We took the personal pronouns of the corpus and looked at what people's guesses were for these pronouns. Looking at what people filled in, we might discover if they were pointing to one of the discourse entities in the text, and whether they used a pronoun.

The Annotation of the Corpus

In order to estimate $P(\textit{referent})$, the corpus needed a coreference annotation. We used the SACR tool (Oberle, 2018), to annotate all texts in which a personal pronoun occurred. This led to the identification of 65 personal anaphoric pronouns. As in other experiments in this thesis, we looked at the following pronouns: *he, him, himself, she, her, herself, they, them, themselves, it, itself*. We excluded cataphora, deictic pronouns and possessives.

TABLE 5.2: An example of one text in the Provo Corpus annotated with coreference chains. The chain with multiple referents is annotated in red, and the other singleton chains are displayed in gray. The identifiers of the chains are at the beginning of the boxes. Note that the number of the identifier does not necessarily indicate the order of appearance in the text. For one of the mentions, *he* encircled in red, we included cloze task data in the table under the text. We can read in this table what the participants guessed at the place were the word *he* occurred in the text. We annotated the guesses with the identifier of the coreference chains to which the guessed word would have belonged and with *UNK* when it was unclear if the guess refered to an entity or whether it was non-referntial. For each guess it is specified how many participants made this guess in the column *counts*.



ID	word in text	guess	entity	counts
QID786	he	to	UNK	23
QID786	he	that	UNK	7
QID786	he	he	M4	3
QID786	he	never	UNK	3
QID786	he	himself	M4	1

To annotate the cloze task data, for each guess, we estimated with which coreference chain the guess was compatible. The guess was than annotated with the identifier of this chain (obtained by the SACR tool), or with the label 'UNK', if it could not be determined whether the guess referred to a discourse entity. We used the label 'new' if it was clear that the participant referred to a new discourse entity and 'poss' if the word was a possessive. An example of this annotation can be found in Table 5.2.

Estimation of Probabilities

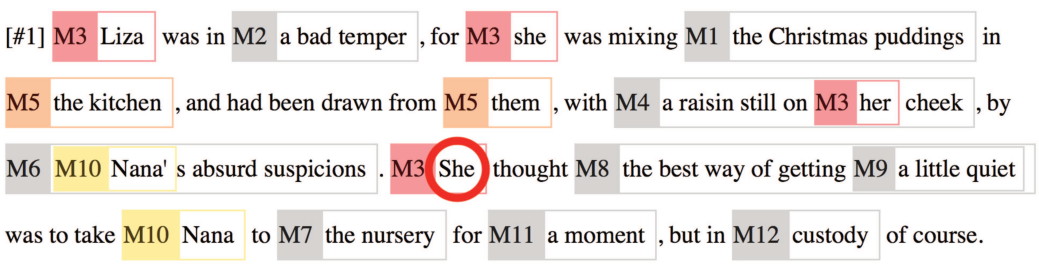
For all the anaphoric pronouns in the corpus, we calculated $P(pronoun|referent)$ and $P(referent)$, using the annotation of the guesses. We also estimated the probability that the participants guessed the exact form of reference, this probability will be referred to by $P(exact\ form)$. We used this probability, to also measure the influence of lexical prediction. Hereunder we describe how the probabilities were calculated and an example of calculation for the pronoun *she* is given in Table 5.3.

- $P(pronoun|referent) = \frac{number\ guesses\ that\ are\ pronouns\ and\ point\ at\ the\ referent}{number\ of\ guesses\ that\ point\ at\ the\ referent}$

- $P(referent) = \frac{number\ of\ guesses\ that\ point\ at\ the\ referent}{number\ of\ guesses\ that\ are\ different\ from\ UNK}$
- $P(exact\ form) = \frac{number\ of\ guesses\ that\ are\ exactly\ the\ word\ in\ the\ text}{number\ of\ guesses}$

TABLE 5.3: This is an example from our corpus to illustrate how we calculate the probabilities $P(exact\ form)$, $P(referent)$ and $P(pronoun|referent)$ using the cloze task data. The guesses humans made for the pronoun surrounded in red are given in the table. We see that out of 42 participants, 12 guessed correctly *she*. Therefore, the probability $P(exact\ form) = \frac{12}{42}$. This probability is used to calculate the lexical prediction probability. For the probability $P(referent)$, we looked at all the answers that correspond to any referent and counted how many times *M3*, the right referent, was mentioned. We devided this number by the total number of answers that were different from UNK. We counted in total 27 non-UNK answers and 20 of them are mentions of the referent, so $P(referent) = \frac{20}{27}$. For the probability $P(pronoun|referent)$, we looked at all the references made to the entity *M3*, that is 20, and looked at the portion that are pronouns, in this case 12 out of 20. Thus, $P(pronoun|referent) = \frac{12}{20} = 0.6$.

ID	word in text	guess	entity	counts
QID207	she	she	M3	12
QID207	she	liza	M3	8
QID207	she	nana	M10	4
QID207	she	this	UNK	2
QID207	she	when	UNK	2
QID207	she	burglars	new	1
QID207	she	by	UNK	1
QID207	she	her	poss	1
QID207	she	however	UNK	1
QID207	she	luckily	UNK	1
QID207	she	nevertheless	UNK	1
QID207	she	of	UNK	1
QID207	she	so	UNK	1
QID207	she	still	UNK	1
QID207	she	that	UNK	1
QID207	she	the	UNK	1
QID207	she	then	UNK	1
QID207	she	we	new	1
QID207	she	what	UNK	1



Note that for the calculation of $P(\textit{referent})$, all the occurrences that are labels *UNK* are excluded. We are aware that this is a problem, because *UNK* is a very frequently occurring label. Therefore, the estimations of $P(\textit{pronoun}|\textit{referent})$ and $P(\textit{referent})$ are probably very noisy because out of the around 40 guesses per pronoun, a lot of data has to be discarded. The data from the cloze task is therefore only partly usable. In the perspective section of this Chapter, Section 5.6, we discuss how the estimation of probabilities based on human data can be enhanced.

5.3.3 Reading Times and Zones

For every word in the Provo Corpus, different reading times and metrics are given. That is to say, every word in the corpus is considered as a reading zone. This means that information about individual fixations are not provided and that zones other than words cannot be defined.

We decided to study the region of the pronoun. Because the boolean reading metric that says whether a pronoun is fixated or not, presented in Section 4.7.3, yielded results in the experiments presented in Chapter 4, we decided to include the reading metric *skipping rate*: a boolean that states whether the pronoun was skipped or not.[2] We also included *first fixation* time, because we think that expectation plays an important role in the probabilities related to pronoun resolution we tested on the Provo Corpus. Indeed, first fixation is a metric of very early processing.

In addition to reading metrics for early processing, we also included two metrics of later processing, to check whether the effects do also show in later processing. We used the regression path reading time and the total number of fixations (see Section 2.3.2 for an explanation about these reading times).

5.3.4 Statistical Model

To investigate the influence of the pronoun resolution probabilities on the reading of the pronouns in the Provo Corpus, we used generalized linear models with random factor structures estimated in a Baysian probability framework implemented by the package brms (Bürkner, 2017) in R (R Development Core Team, 2008). A description of this method can be found in Section 4.7.4.

All models contained random intercepts and slopes for participants and pronoun instances. We also included two control factors: word length and the position of the pronoun in the sentence. All factors, except for the reading times and the skipping rate were scaled (see Section 2.3.3 for an explanation about scaling). The distributions we choose for the output data were adapted to the type of reading metric. An ex-gaussian distribution was chosen for the first fixation duration and the regression path reading time. This skewed distribution is known in the psycholinguistic literature to be a good fit for reading time data (e.g. Staub, 2011). We used a bernoulli distribution for the skipping rate and a poisson distribution for the total count of fixations. All models were run with weakly informative priors.

[2] The skipping rate encodes thus the same information as the boolean fixation metric, except that the skipping rate takes the value of *true* when the pronoun is not fixated and the boolean fixation metric attributes the value of *true* when the pronoun is read.

5.4 Results

The results for the Bayesian model of Kehler and Rohde were positive for the first fixation duration and negative for the other three metrics (see Appendix B.1). The model of the first fixation reading times, which has a 95% credible interval between -2.97 and -0.04, indicates that a higher probability $P(\textit{referent}|\textit{pronoun})$, calculated by the Bayesian formula, leads to a shorter reading time. For the other metrics, we could not observe any tendencies. The positive outcome of the first fixation duration was surprising, giving the noisy estimation of the probabilities from the Provo Corpus annotation. We believe that more research is necessary to confirm this result. It should be interpreted with care, also because the two control factors do not seem to contribute to the model.

As for the probability of the referent, we did not find any influence of it on the reading data (see Appendix B.2). It is important to note that these probabilities were not adequately estimated on corpus data, making the negative results no surprise.

On the other hand, the predictability $P(\textit{exact form})$, estimated via the cloze task data, had a small effect on reading time. This has been demonstrated by the skipping rate and the first fixation duration. For the skipping rate, the 95% credible interval lied between -0.53 and -0.01 for $P(\textit{exact form})$, whereas for the first fixation duration, it lied between -2.50 and 0.34 (a tendency). These results suggest that the probability of giving the right answer at the cloze task, facilitates the reading of the pronoun, leading to a lower first fixation duration and a higher skipping rate. In other words, the predictability of the word form facilitates the reading. However, $P(\textit{exact form})$, showed to be of no influence on the other two reading time metrics: the total number of fixations and the regression path duration. The models can be found in the Appendix B.3.

5.5 Discussion

We observed that the pronoun resolution probability from Kehler and Rohde's theory and the probability of guessing the exact word form are predictors of first fixation reading time. Guessing the exact word form also seems to have some influence on the skipping rate but the result is not 100% conclusive. On the other hand, none of the probabilities we tested was of influence on the regression path duration, nor the total reading time. We think that the difference between the first fixation duration and the skipping rate on the one side and the regression path duration and the total number of fixations on the other side, is not contradicting. In fact, first fixation duration and the skipping rate reflect very early processing, whereas the regression path duration and the total numbers of fixations also includes instances of later processing. Indeed, it makes sense that the predictability especially influences the very early processing, and not the later. Predictability also plays a role in the model of Kehler and Rohde: the positive result for the first fixation reading time could be caused by this. But, as the formula is composed by various components, it might also be necessary to explore the word after the pronoun as well as other reading time metrics. Effects may appear in a later stadium of processing and occur as spill-overs. This needs to be investigated further in follow-up experiments.

It also has to be investigated whether all components in the Bayesian formula have their role to play. That is to say, the $P(\textit{referent}|\textit{pronoun})$ calculated in the end

must yield better results than $P(pronoun|referent)$ and $P(referent)$ separately and the effect must go beyond lexical prediction: the simple $P(exact\ form)$ probability.

5.6 Perspectives

In this section, we discuss future experiments that we plan to do in order to investigate further whether the Bayesian Theory of pronoun resolution makes correct predictions about processing. We first discuss how the estimation of probabilities can be enhanced. First how $P(pronoun|referent)$ and $P(referent)$ can be estimated more accurately using human responses and then how they can be estimated using algorithms that have been proposed in the field of natural language processing while respecting the idea that $P(pronoun|referent)$ can be estimated using Centering Theory and $P(referent)$ using Coherence.

5.6.1 Estimation of Probabilities from Human Answers

We used the cloze task data provided with the Provo Corpus but we had to discard a large part of the data because much of participant's guesses does either not correspond to discourse referents or it is not clear which discourse referent was intended. Indeed, at places where originally in the text a pronoun appeared, participants often guess *this* or *the*. There is no way we can find out what they intended to say because only the first word they typed in was maintained in the experiment. Therefore, to estimate adequately $P(pronoun|referent)$ and $P(referent)$ it is necessary to collect another type of human responses than just let them guess the next word (cloze task).

An option we want to explore to obtain better probabilities for $P(referent)$ is the *referent prediction task* (Modi et al., 2017), or *referent guessing game* (Tily and Piantadosi, 2009; Kravtchenko, 2014). Like a cloze task, this task is about guessing upcoming material in the text. But an important difference is that not the word-form must be guessed but which referent is mentioned. The most convenient way to obtain enough guesses to build a probability distribution is to recruit participants via crowd-sourcing platforms.

For the referent prediction task, participants have to guess which entity is coming up next in a text. First, the text is shown up to the first noun phrase — without including it — and participants have to guess what would be mentioned next. Then, the text until the second noun phrase is shown and they have to guess again. This goes on until the participants have read the whole text and guessed all the NPs. To choose a referent, participants can click on previous mentions in the text if they think that they are rementioned where the text ends. They can also choose to press a button that says 'new' if they think that a new discourse referent is introduced.

For measuring $P(pronoun|referent)$, we can use the method of Tily and Piantadosi (2009) to measure the probability of using a given type of referential expression (pronoun, name, descriptive noun phrase). They suggest that in a text at places where NPs occur, gaps can be left instead. Participants have to fill in the missing noun phrase by typing in their answer. They can use the referential form they think fits the best in the context. Contrary to the referent prediction task, a part of the context after the missing noun phrase must be shown. This part of the context can be important for resolution. As the referent should be known in $P(pronoun|referent)$, it is important to enable participants to resolve the gap. As with the *referent prediction*

task, this gap-filling task would also be more time and cost efficient if it is done as a crowd-sourcing study.

5.6.2 Estimation of Probabilities with NLP-Tools

Only after it has been shown on probabilities provided by humans that the Bayesian model of Kehler and Rohde (2013) can predict reading behaviour on natural texts, it is worth the effort of estimating the parameters of this model by computational methods. It would be interesting to compare the results of the probabilities estimated on human data to probabilities estimated by computational models. Using automatically estimated probabilities has advantages over probabilities estimated on human data. First, it is less time and money consuming to estimate the probabilities on a new resource. Whereas for every new text a new crowd-sourcing experiment must be set-up, tools that estimate the probabilities automatically are easier to apply on new texts. Second, it enables us to investigate the theory of Kehler and Rohde (2013) into more depth: with human probabilities, only the relevance of $P(pronoun|referent)$ and $P(referent)$ and their Bayesian combination can be estimated. But the Bayesian Theory also states that $P(pronoun|referent)$ is determined by Centering and $P(referent)$ by Coherence. Automatic systems that have implemented Centering and Coherence could therefore evaluate whether not only the Bayesian formula is correct, but also whether the underlying processes can be simulated.

In the following paragraphs, we discuss which automatic systems can be used to estimate $P(pronoun|referent)$ and $P(referent)$ for the Bayesian model of pronoun resolution.

Estimation of P(pronoun|referent)

For $P(pronoun|referent)$ we thought about using the entity grid approach (Barzilay and Lapata, 2008). The entity grid is a framework developed in order to make automatically generated texts more coherent by respecting the structure of entity chains. It is inspired by theories based on the notion of *saliency* such as Centering (Grosz, Joshi, and Weinstein, 1983; Grosz, Weinstein, and Joshi, 1995).

The entity grid algorithm captures the structure of reference in a text by assessing what the transitions between the referential expressions are. For a text, all referents are listed as the columns of a grid. The lines of the grid are the sentences of a text. Then, for every sentence, the cells of the grid can be filled in: whether a mention appeared in the sentence or not. This grid can then be used to calculate transitional probabilities on referents. For example: what is the probability that a referent is mentioned again, given that it was mentioned in the last two sentences? More crucially, entities grids are not only a means to measure in which sentence referents occurred, but also in which form. In their article, Barzilay and Lapata (2008) show how the grid can be filled in considering the syntactic functions of the referents. Using the grid, questions can be answered such as: how likely is it that a referent that is the syntactic object is mentioned as a subject in the next sentence? Barzilay and Lapata (2008) illustrated this with the examples given in Figure 5.1.

The syntactic function is not the only feature on which transitional probability is calculated. The final model of Barzilay and Lapata (2008) also includes a frequency feature: a distinction is made between infrequent and frequent entities. It is easily

imaginable how the entity grid can be adapted to include information about the referents being mentioned as pronouns or not. Just as the syntactic function and the frequency factor, it can be added to the entity grid. Then, this grid can estimate $P(pronoun|referent)$. For our implementation of the entity grid, we are considering the Brown Coherence Toolkit, that has a well-documented off the shelve implementation (Elsner and Charniak, 2008; Elsner and Charniak, 2011) [3].

[3] https://bitbucket.org/melsner/browncoherence/overview

FIGURE 5.1: An illustration from the article of Barzilay and Lapata (2008) of how the entity grid approach can provide a probability distribution on entity transitions in a text. In the first subfigure we see a text that is annotated for the syntactic functions of subject, object and other. In the second subfigure, it is captured in an entity grid in which syntactic function entities are mentioned in the sentences of a text. In the third subfigure, the probability distribution for the bi-gram transitions of syntactic functions in two succeeding sentences are given.

(A) An example of a text.

Summary augmented with syntactic annotations for grid computation.

1 [The Justice Department]$_S$ is conducting an [anti-trust trial]$_O$ against [Microsoft Corp.]$_X$ with [evidence]$_X$ that [the company]$_S$ is increasingly attempting to crush [competitors]$_O$.
2 [Microsoft]$_O$ is accused of trying to forcefully buy into [markets]$_X$ where [its own products]$_S$ are not competitive enough to unseat [established brands]$_O$.
3 [The case]$_S$ revolves around [evidence]$_O$ of [Microsoft]$_S$ aggressively pressuring [Netscape]$_O$ into merging [browser software]$_O$.
4 [Microsoft]$_S$ claims [its tactics]$_S$ are commonplace and good economically.
5 [The government]$_S$ may file [a civil suit]$_O$ ruling that [conspiracy]$_S$ to curb [competition]$_O$ through [collusion]$_X$ is [a violation of the Sherman Act]$_O$.
6 [Microsoft]$_S$ continues to show [increased earnings]$_O$ despite [the trial]$_X$.

(B) The entity grid for the text in Figure 5.1a.

A fragment of the entity grid. Noun phrases are represented by their head nouns. Grid cells correspond to grammatical roles: subjects (S), objects (O), or neither (X).

	Department	Trial	Microsoft	Evidence	Competitors	Markets	Products	Brands	Case	Netscape	Software	Tactics	Government	Suit	Earnings	
1	S	O	S	X	O	–	–	–	–	–	–	–	–	–	–	1
2	–	–	O	–	–	X	S	O	–	–	–	–	–	–	–	2
3	–	–	S	O	–	–	–	–	S	O	O	–	–	–	–	3
4	–	–	S	–	–	–	–	–	–	–	–	S	–	–	–	4
5	–	–	–	–	–	–	–	–	–	–	–	–	S	O	–	5
6	–	X	S	–	–	–	–	–	–	–	–	–	–	–	O	6

(C) The transitional bi-gram probability based on the grid in Figure 5.1b.

Example of a feature-vector document representation using all transitions of length two given syntactic categories S, O, X, and –.

	SS	SO	SX	S–	OS	OO	OX	O–	XS	XO	XX	X–	–S	–O	–X	––
d_1	.01	.01	0	.08	.01	0	0	.09	0	0	0	.03	.05	.07	.03	.59

Estimation of P(referent)

For the estimation of $P(referent)$ we want to use the approach of Modi et al. (2017). Modi et al. (2017) developed a computational model for the *referent prediction task*. In the referent prediction task, a system has to predict which referent will be mentioned next in a text based on the left context. Modi et al.'s model is based on so-called *script knowledge*. Script knowledge is knowledge about typical scenarios, such as taking a bath, or ordering food in a restaurant. It can be used for artificial intelligence tasks. The model is trained, developed and tested on a corpus, called the InScript Corpus (Modi et al., 2016), that contains about 1 000 stories centred around ten scenarios. The system was evaluated on the task of entity prediction with different feature sets and compared to human guesses recorded on the InScript Corpus via Amazon's Mechanical Turk: a commonly used crowd-sourcing platform. A model that only included linguistic features, such as the grammatical function, performed with an accuracy of 49,53%. This model, augmented with script knowledge, got an improved score of 62,63% accuracy. Humans performed this task with 73,63% accuracy.

We are considering to use this model with the linguistic features only, because our eye-tracking corpora are not annotated with script knowledge. However, short after the publication of Modi et al. (2017)'s model, a new article (Ji et al., 2017) reported to obtain better scores on the referent prediction task, without needing script knowledge. Ji et al. (2017) proposed a neural network language model that has integrated knowledge of entities. It is reported to perform at human level with an accuracy of 74,23% on the data of the InScript Corpus. Therefore, in addition to Modi et al. (2017)'s model, we also consider this model.

The challenge with the models of Modi et al. (2017) and Ji et al. (2017) is to use the programming code corpora on which the code was not developed. Indeed, it has to be investigated whether the code of Modi et al. (2017) — that is available on request — can be run easily. The code of Ji et al. (2017) is available on-line but documentation and a trained version of the model are lacking. It has to be investigated whether the models can be exploited and otherwise, how we can reimplement them.

5.6.3 Other Corpora

We used the Provo Corpus because the number of participants in the eye-tracking data is high and it provides a fair number of pronouns as well. Moreover, cloze task data for this corpus was collected and we exploited it to estimate $P(pronoun|referent)$ and $P(referent)$. However, because of some concerns with the corpus, the question arises whether a future experiment should also be done using this resource or whether we have to consider another resource.

The first problem we can cite is the choice of texts: some texts fragments are from very well known stories, such as *Peter Pan* and *the Wizard of Oz*. As we have observed in cloze task data, this has an influence on the predictability of the texts: it seems that participants are familiar with the text. This becomes even more clear for a number of somewhat ill-chosen texts in which anaphoric pronouns figure without any antecedent (the beginning of the story was cut-off). Despite of the absence of an antecedent in the context, participants guess the personage, because they are familiar with the story. See Figure 5.2 for an illustration.

FIGURE 5.2: A text from the Provo Corpus that has anaphoric pronouns but of which the antecedent was not included in the fragment. The pronoun *He* (circled in red) was often guessed as *Captain*, or *Hook* by the participants in the cloze task. Therefore we conclude that the participants recognized the text as from *Peter Pan*.

[#1] Seeing Peter slowly advancing upon him through the air with dagger poised, he sprang upon the bulwarks to cast himself into the sea. He did not know that the crocodile was waiting for him; for we purposely stopped the clock that this knowledge might be spared him: a little mark of respect from us at the end.

Whereas this problem could be resolved by eliminating the problematic text fragments, there is a second problem that seems to be more difficult to resolve: there is no access to all the data. Even if the availability of the Provo Corpus is rather good and we appreciate the efforts made by the authors to publicly release the resource, no other reading metrics can be used than those provided. The zone of interest is always one word long because only pre-calculated reading times are available, not the data of all fixations.

The two problems we discussed are not a reason to just dump the corpus but we should keep in mind that it would be a good idea to do tests on different type of corpora as well. Of course, if probabilities can be estimated automatically with the help of NLP-tools, nothing prevents us of including also data from the Dundee Corpus in a future study. Another option is to include the Natural Story Corpus (see Section 4.4.3). This is a corpus with data from a self-paced reading experiment. An advantage of this corpus would be that the numbers of participants and anaphora is probably high, which would give statistical power to the study. A downside is that the corpus is full of many low-frequency grammatical constructions. The sentences look natural, but are actually hard to process, even though this was done on purpose because the authors of the corpus are particularly interested in these syntactic phenomena. We do not know how this feature of the corpus affects pronoun processing and therefore argue that it would be the best to validate the Bayesian Theory on different corpora. A last solution would be to collect a new resource ourselves. But we have to keep in mind that this would be very costly and this resource would also have some downsides, just like any resource.

5.7 Conclusion

In this chapter, we studied the Bayesian model of pronoun interpretation proposed by Kehler and Rohde (2013). The Bayesian model states that pronoun interpretation can be modelled as $P(referent|pronoun)$ and that it can therefore be rewritten with Bayes' Theorem. If this is done, two terms, $P(pronoun|referent)$ and $P(referent)$, need to be estimated in order to calculate $P(referent|pronoun)$. According to Kehler and Rohde (2013), $P(pronoun|referent)$ is the problem of pronoun production and it is determined by saliency factors. They argue that it should be estimated by looking at

theories such as Centering Theory (Grosz, Weinstein, and Joshi, 1995) that state that the occurrence of pronouns depends on text structure. $P(referent)$ is determined by world knowledge. Kehler and Rohde (2013) argue that the term can be estimated efficiently if the *Coherence* framework is used. This framework suggests that $P(referent)$ is determined by reasoning on the context.

We wanted to test this model because it proposes a formula based on Bayes' theorem to calculate the probabilities of pronoun interpretation. The Bayesian formula gives us a concrete framework: it makes the theory measurable. In addition, the theory specifies the influence of multiple factors, which is in line with the psycholinguistic literature on pronoun resolution that demonstrated that many factors are of influence on pronoun interpretation. And last but not least, it supposes a probability space over possible referents of a pronoun. Therefore, it is also compatible with our hypothesis that competition plays a role in pronoun resolution that we presented in Section 4.5.

We proposed an experiment to see whether corpus data provide evidence for the claims of the theory because evidence from corpora would be a great argument in favour of the robustness of the theory. There are two questions that need to be answered concerning the Bayesian model. The first is: is the Bayesian proposal correct? That is to say: is it true that pronoun interpretation is determined by pronoun production and next mention biases? The second question — that is hard to answer if the answer to the first question turns proves negative — is: is $P(pronoun|referent)$ indeed determined by Centering and $P(referent)$ by Coherence?

In this chapter, we only provided a pilot study for the first question. We used the Provo Corpus (Luke and Christianson, 2016): an eye-tracking corpus that also provides cloze task data. We used the cloze task data to approximate the two probabilities ($P(pronoun|referent)$ and $P(referent)$) and tested on the reading data whether $P(referent|pronoun)$ calculated by the Bayesian formula could predict reading behaviour of participants. We also checked whether the probability of predicting the exact form, $P(exact\,form)$, and the probability of mentioning a referent, $P(referent)$, could be predictors.

We found a positive result for the Bayesian model and for the probability of predictability of the exact form. However, the method we used to estimate the probabilities was not precise and therefore our results needed to be interpreted with caution. We therefore conclude that our results are encouraging but that we need more sophisticated methods to improve the estimation of the probabilities. In the perspectives section, we discussed how our study could be improved.

A first improvement is to get better estimations of the probabilities $P(referent)$ and $P(pronoun|referent)$ based on human data. The cloze task data from the Provo Corpus was too noisy to get accurate estimations of these probabilities. We therefore propose to conduct new crowd-sourcing experiments that are better at estimating the necessary probabilities. A second improvement that could help to answer the question whether $P(pronoun|referent)$ is indeed determined by Centering factors and $P(referent)$ by coherence factors, is to use computational models that implement these Centering and Coherence theories to estimate the probabilities. In the natural language processing literature, we found interesting options we want to pursue in future work. A final improvement would be to include also other corpora with reading times, notably the Dundee Corpus or the Natural Story Corpus. As each corpus has its own biases, we think it is wise to repeat experiments on multiple resources.

If the Bayesian model would yield results in future experiments in which we could also investigate the nature of the factors that contribute to $P(pronoun|referent)$ and $P(referent)$, it would be an important finding for the NLP-community. Indeed, today's resolution system are still very much focussed on saliency factors and not much on semantic knowledge. Also, it would be interesting to see whether the role of pronoun production in the process of pronoun interpretation that is specified by the Bayesian model is relevant for automatic systems.

Chapter 6

Conclusion and Perspectives

We presented various cognitive computational models of pronoun resolution in this thesis. The absence of literature about this topic resulted in a lot of exploratory work. We gained a better understanding of the methods to explore reading time data for the computational modelling of pronoun resolution. We investigated the possibility of using existing pronoun resolution systems from the NLP-community for the purpose of cognitive computational modelling. We explored how Information Theory can inspire models of pronoun resolution and we made a start in the evaluation of a recent psycholinguistic theory of pronoun resolution (Kehler and Rohde, 2013). In this chapter we will reflect on what we learned from the experiments presented in this thesis and what the perspectives for our research are.

6.1 Summary of Contributions

In Chapter 2, we investigated whether pronoun resolution biases discovered in psycholinguistic studies could also be demonstrated in reading time corpora. We modelled the first pass reading times of the pronouns of the Dundee Corpus (Kennedy, Hill, and Pynte, 2003) with linear mixed effects models. A challenge we faced was that it is not clear where the effect of pronoun resolution shows in the data. The literature suggest that the effect might appear on the pronoun, one word before the pronoun or as a small or large spillover effect. Therefore, we modelled reading times in a window of six words: from one word before the pronoun up to four words after the pronoun. Our results suggest that the effects of pronoun resolution show mainly as spillover effects. However, spillover effects occurred one, two, three, or four words after the pronoun depending on the tested factor. In addition, they did not last over multiple regions. So it remains difficult to say where the effect of pronoun resolution occurs exactly in reading data.

Because we tested on six regions, it would be necessary to apply a statistical correction for multiple testing to the results we obtained. But no effects we found remain significant if we did. We think there is a low probability that all the results we found are in reality false positives. However, there is a good chance that some of them are, and we cannot tell which ones. We think that reading time corpora are an interesting resource to evaluate pronoun resolution models on but that modelling pronoun resolution as a bunch of resolution biases is only possible with a very large dataset to have enough statistical power to support a complex model. Also, the approach does not allow to evaluate theoretical claims beyond the level of individual resolution biases.

In Chapter 3, we simulated human resolution biases attested in psycholinguistic experiments with an automatic pronoun resolver trained on a large corpus. The model chooses antecedents of ambiguous pronouns in a manner very similar to humans. This means that human resolution biases are also reflected as statistical trends in the language. They can thus be learned on corpus by automatic systems. Therefore, we conclude that NLP-systems have a big potential to inspire cognitive computational models. In Chapter 3, we also saw how computational modelling can be used to specify theoretical claims: we saw that by using an automatic parser, we got a better understanding of what grammatical parallelism is.

In Chapter 4, we formulated a cost metric of pronoun resolution based on Information Theory (Shannon, 1948). Our motivation was that for the domain of syntax, concepts from Information Theory allowed to formulate processing hypotheses that were supported by corpus studies. We hypothesized that pronoun resolution is influenced by the level of competition there is amongst antecedent candidates. When there is more competition, pronoun resolution has a higher cognitive load. We used entropy, a measure of ambiguity over a probability distribution, to implement this idea. To estimate the probability distribution over the antecedent candidates of a pronoun, we used probabilistic NLP pronoun resolution systems.

We evaluated the cost metric in two experiments. In the first experiment, we tested whether self-paced reading times recorded during psycholinguistic experiments could be modelled correctly. Most of the reading times were correctly simulated by the cost-metric, but not all. A potential source of error was the somewhat cumbersome manner in which the probability distribution was estimated. For our second experiment, in which we evaluated the entropy cost metric on reading data of the Dundee Corpus, we thus chose a state of the art NLP-system that could estimate the probability distribution in a direct manner (Lee et al., 2017). We used a mixed effects model in a Bayesian statistics framework to test whether more participants fixated a pronoun when there was more competition amongst its antecedent candidates. The use of this reading metric prevented us from testing multiple zones as in the experiment of Chapter 2. The result was positive, so we concluded that competition amongst antecedent candidates of a pronoun is of influence on human pronoun resolution and that more competition leads to more cognitive load. When pronoun resolution is modelled, it is thus necessary to take the competition amongst antecedent candidates into account.

In Chapter 5, we made a start in testing existing psycholinguistic theories on corpus data. We investigated whether we could find evidence from reading time corpora for a theory of Kehler and Rohde (2013) (K&R-Theory). Our choice for this theory was motivated by the fact that it proposes a probabilistic formula to calculate pronoun resolution and that it makes use of a probability space over antecedent candidates. According to K&R-Theory, pronoun resolution can be formulated as the probability of a resolution to a referent given that a pronoun appears in the text: $P(ref|pro)$. A conditional probability can always be decomposed making use of Bayes' Theorem. Therefore, the pronoun resolution problem can be estimated by making use of the probabilities from the Bayesian decomposition. K&R-Theory argues that one of the probabilities can be calculated by saliency factors and the other by applying world knowledge. It supports its claims with evidence from completion studies. We believe that it is useful to also evaluate the theory on more natural data: if the theory makes relevant predictions on corpus, it would be strong evidence in favour of it.

We made a start in evaluating this theory by looking whether we find corpus evidence for the Bayesian proposal. We estimated the parameters of the Bayesian formula by means of cloze task data. Then, we evaluated the model by testing it on four types of reading data from eye-tracking using mixed effect models in a Bayesian statistics framework. We found that the model was a predictor of first fixation reading time of pronouns, but we did not find any other positive results for the remaining three metrics. However, the estimations from the cloze task data were very noisy. Therefore, we cannot see the result as hard evidence for the Bayesian proposal of K&R-Theory. We conclude that the results are encouraging, but that the parameters should be estimated with more accuracy. We proposed to estimate them using human responses from crowd-sourcing platforms on the one hand and computational implementations on the other hand. The crowd-sourcing responses would allow us to evaluate the Bayesian proposition properly. The computational implementations, in their turn, allow to test whether it is indeed saliency and world knowledge that are involved in the pronoun resolution process.

6.2 General Conclusions

Cognitive computational modelling is a means that is complementary to classical psycholinguistic studies. Computational modelling is a tool to specify theories more and obtain better definitions. It allows theories to make quantitative and measurable predictions. Therefore, it is a powerful tool to compare theories but also to measure the size of effects. In addition, computational models can be run on corpus data. This is useful because it allows us to measure the impact of linguistic effects on natural data. It is also very important to the process of pronoun resolution. Pronoun resolution is a phenomenon that is influenced by many linguistic factors and theories attribute a crucial role to discourse structure. Evaluating pronoun resolution models on natural discourse is therefore important and cognitive computational models can be of great help in this process.

Biases that have been attested in psycholinguistic experiments are also present as frequency tendencies in corpus data. For example in English, pronouns are resolved faster by humans when they refer to an antecedent that is the syntactic subject and at the same time, more pronouns refer to antecedents in the subject position. This leads us to conclude that human language processing is sensitive to statistical tendencies.

A second finding in our thesis that leads to the same conclusion is that information theoretical cost metrics make accurate predictions about language processing. Information Theory provides tools to measure the quantity of information conveyed by linguistic structures, on condition that these linguistic structures are modelled as probabilistic events. Surprisal and entropy of linguistic structures have been shown accurate predictors of cognitive load experienced by humans in the psycholinguistic literature and also in this thesis.

6.3 Perspectives

The perspectives on the short term are to complete and repeat our experiments. In Section 5.6, we discussed how we count to continue to work on the model of Kehler and Rohde: we need to estimate its parameters more accurately. We plan to use human answers from a crowd-sourcing task to estimate the parameters of mentioning

referents and pronoun production. We also want to evaluate the claims of K&R-Theory that pronoun production is related to saliency and Centering and mentioning referents to world-knowledge and Coherence. We believe that providing computational implementations for these claims is very important. In our opinion, what saliency and coherence is exactly is left for a large part unspecified in K&R-Theory. Some examples are given of manipulations that make some referents more salient then others, such as being in the subject position, but no full explanation of how to measure saliency and world knowledge is provided. Computational models can be used to give better definitions to these concepts.

To test K&R-Theory, we discussed the possibility to incorporate other reading time corpora into the study. These other reading time corpora could also be interesting to other experiments presented in this thesis, in particular for testing our information theoretical cost metric. For reading time data, there are still two interesting corpora for English the we have not explored yet: the Natural Story Corpus (Futrell et al., 2018) and the Geco Corpus (Cop et al., 2017).

The Natural Story Corpus contains data of self-paced reading.[1] It would be interesting to compare eye-tracking data to self-paced reading data: in self-paced reading data effects of pronoun resolution might be more local than in eye-tracking data because participants cannot skip words or go back in the text. The texts in the Natural Story Corpus have the particularity that they were designed, or rather adapted, to contain a large number of rare syntactic constructions. Therefore, the authors of the corpus estimate that the difficulty of the reading can be higher than on average. We find this fact very intriguing and are curious whether it changes pronoun resolution. However, we realize that a corpus that is particularly hard to read is not the best starting point for a study of pronoun resolution in natural text reading.

The Geco Corpus is based on a novel, so it allows us to study whether there are differences between the literature genre and genres of other corpora, such as the news genre in the Dundee Corpus. Moreover, it contains reading times of both English mono-linguals and English-Dutch bi-linguals. Using this data source would thus enable us to make cross-lingual comparisons. Of course, to study pronoun resolution in natural data cross-linguistically, we can also include eye-tracking data from mono-lingual corpora of other languages than English, such as the French Treebank (Abeillé, Clément, and Toussenel, 2003). But the specificity of the Geco Corpus is that exactly the same discourse for English and Dutch is available.

Beyond the possibility to conduct new experiments on reading time data, we could extend our research to other types of data. We can for example try to replicate the results of the entropy cost metric on fMRI data. fMRI data of participants listening to a chapter of *Alice in Wonderland* has been made available by Brennan et al. (2016). Finally, we could also consider building our own corpus but we believe that we must first study the already available resources. We believe that a clear picture will emerge when more and more resources are studied, so we consider constructing our own resource as one of the possibilities to go forward but not as a final solution.

Our long-term perspectives focus on the one hand on finding a better account for the cognitive processes involved in resolution and on the other hand on integrating anaphora resolution more into a broad discourse and syntax framework. We elaborate these ideas in the last paragraphs of the thesis.

[1] Participants read a text word by word. Only one word appears on the screen at the time and they have to push a button — for example the space bar — to read the next word. Reading times are recorded for every word.

Our work about pronoun resolution from a cognitive perspective by using NLP-programs for coreference resolution gives rise to questions concerning the nature of pronoun and coreference resolution. On one hand, in psycholinguistics, pronoun resolution comes down to resolving a pronoun to an antecedent. In most psycholinguistic experiments, coreference chains are very small, only containing one or two mentions. Therefore, the cognitive status of coreference chains is most often not addressed. On the other hand, in the NLP-community, it is often stated that it is important to attach referential expressions not only to a single antecedent but to the whole coreference chain. However, the field has trouble to model these coreference chains accurately: systems that perform coreference resolution on a mention to mention basis yield better results with less efforts. This begs the question what the status of coreference chains is in the human mind. Do humans resolve pronouns to a previous mention, or to an abstract entity or to both? Finding an answer to these questions in future work would be of value to both the field of psycholinguistics and the NLP-community.

Another question that we would like to answer on the long term is whether pronoun resolution and other types of coreference resolution have similar underlying mechanisms. Often, NLP-models do not distinguish between pronouns and other forms of coreference.[2] But the use of a pronoun — a very short form with almost no semantic information — has a special function. So, it is quite likely that the underlying cognitive processes of pronoun and coreference resolution also show differences. Studying these differences would help us to better understand anaphora and coreference in general.

Another direction of study that we want to pursue is to investigate further the role prediction plays in pronoun resolution. In Chapter 4, we modelled the competition amongst antecedent candidates with a cost metric based on entropy. However, we are also interested in the question how a metric of prediction, such as surprisal, is of influence on pronoun resolution. Indeed, we would like to determine which part of the cognitive load is determined by deceived expectations, or relieved by correct predictions, and which part is due to a calculation of what the right referent of the pronoun is, once the pronoun has been encountered. In literature about pronoun resolution and in every NLP-system, resolution is often modelled as a calculation that is effected the moment that a pronoun is encountered. But as K&R-Theory points out, anticipation can also play a role. Moreover, participants could anticipate on different levels: anticipations could be of lexical or semantic nature. In future work, we would like to determine what the role of prediction in the resolution process is and investigate what happens when anticipations are confronted to evidence for resolution to one specific antecedent. Modelling these two steps in the resolution process separately could also lead to new approaches in NLP.

In addition to studying the cognitive mechanisms involved in pronoun resolution in more detail, we also aim at a better integration of discourse structure into the modelling of anaphora resolution. Discourse structure is assumed to be of importance to anaphora resolution. Often, the influence of discourse structure is addressed by factors that are more or less an approximation of discourse structure. Many of such factors can be found amongst the resolution biases attested in psycholinguistic literature: syntactic function, order of mention, or the distance between the pronoun

[2]However, today it is common to use neural networks for coreference resolution and it could be the case that distinctions between pronominal anaphora and other types of coreference are learned automatically by the networks.

and its antecedent. Using this type of factors results into a 'flat' representation of discourse and does not take into account coherence relations. We want to investigate how a more hierarchical representation of discourse can be implemented in computational models of anaphora resolution. To do so, we have to study formal theories of discourse and see how pronoun resolution systems can be integrated into their implementations while remaining robust and of large coverage. We believe that enriching anaphora resolution systems with a more sophisticated representation of discourse will give a boost to coreference resolution research: since the algorithm of Lappin and Leass (1994) the type of features used in automatic anaphora resolution systems have not changed much. What changed were the scoring functions of the algorithms — supported by developments in machine learning — the availability of corpora and the modelling of the anaphora resolution problem as a coreference task. We believe that it is now time to address the modelling of discourse structure.

A last perspective of this thesis is the integration of pronoun and anaphora resolution in a broader framework of cognitive computational modelling. We primarily want to obtain a better integration with syntax. As discussed in Chapter 4, cognitive computational modelling of syntax is a field that develops quickly, especially by making use of information theoretical cost metrics. In our thesis, we concluded that Information Theory also allows to formulate accurate hypotheses about pronoun resolution. Therefore, it is thinkable to integrate both syntax and pronoun resolution into one cognitive computational model based on information theoretical assumptions. This is an idea that has been discussed in the literature, for example in the work of Dubey, Keller, and Sturt (2013) (see Section 4.4.1) where an attempt was made to model cognitive cost of syntax and discourse simultaneously. However, both the syntactic model and the discourse model were estimated by simple heuristics. It is worthwhile to build broad coverage models of linguistic processes including influences from various linguistic levels. We want to continue this research by using more fine-grained models.

Appendix A

Bayesian Model from Chapter 4

Hereunder, we present the summary of the model object as provided by R in which we tested the cost metric of relative entropy discussed in Section 4.7.5. The factor *rel_ent* is the relative entropy metric. We see that the 95% credible interval does not cross 0.

```
 Family: bernoulli
  Links: mu = logit
Formula: fixated ~ length_in_chars + frequency_bnc + comma + hard_punct +
    rel_ent + (1 + rel_ent | participant) + (1 | dundee_tokens)
   Data: pro (Number of observations: 11040)
Samples: 6 chains, each with iter = 3000; warmup = 1500; thin = 1;
         total post-warmup samples = 9000
    ICs: LOO = NA; WAIC = NA; R2 = NA

Group-Level Effects:
~dundee_tokens (Number of levels: 1104)
              Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)     0.51      0.03     0.44     0.58       3515 1.00

~participant (Number of levels: 10)
                   Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)          0.42      0.12     0.25     0.71       3402 1.00
sd(rel_ent)            0.03      0.02     0.00     0.09       5827 1.00
cor(Intercept,rel_ent) 0.03      0.43    -0.78     0.80       9000 1.00

Population-Level Effects:
                Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept          -0.23      0.14    -0.51     0.05       2782 1.00
length_in_chars     0.41      0.04     0.33     0.49       9000 1.00
frequency_bnc      -0.19      0.04    -0.27    -0.10       9000 1.00
comma TRUE          0.58      0.17     0.25     0.91       9000 1.00
hard_punct TRUE     0.04      0.11    -0.18     0.25       9000 1.00
rel_ent            -0.07      0.03    -0.13    -0.01       9000 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Appendix B

Bayesian Models from Chapter 5

B.1 Model's output for P(referent|pronoun)

B.1.1 P(referent|pronoun) Skipping Rate

```
 Family: bernoulli
  Links: mu = logit
Formula: IA_SKIP ~ Word_Length + Word_In_Sentence_Number + (1 + p_kr | Participant_ID)
+ (1 + p_kr | Word_Unique_ID) + p_kr
   Data: eye_tracking (Number of observations: 2462)
Samples: 6 chains, each with iter = 3000; warmup = 1500; thin = 1;
         total post-warmup samples = 9000
    ICs: LOO = NA; WAIC = NA; R2 = NA

Group-Level Effects:
~Participant_ID (Number of levels: 84)
                    Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)           0.77      0.09     0.60     0.97       3332 1.00
sd(p_kr)                0.12      0.08     0.01     0.31       2856 1.00
cor(Intercept,p_kr)    -0.17      0.38    -0.81     0.65       9000 1.00

~Word_Unique_ID (Number of levels: 62)
                    Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)           0.76      0.10     0.58     0.98       3317 1.00
sd(p_kr)                0.17      0.13     0.01     0.47       1950 1.00
cor(Intercept,p_kr)     0.18      0.45    -0.71     0.89       5251 1.00

Population-Level Effects:
                        Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept                  -1.56      0.15    -1.85    -1.28       3301 1.00
Word_Length                -0.28      0.13    -0.54    -0.03       3450 1.00
Word_In_Sentence_Number    -0.29      0.12    -0.53    -0.05       3630 1.00
p_kr                       -0.10      0.12    -0.35     0.14       3844 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

B.1.2 P(referent|pronoun) First Fixation Duration

```
Family: exgaussian
  Links: mu = identity; sigma = identity; beta = identity
Formula: IA_FIRST_FIXATION_DURATION ~ Word_Length + Word_In_Sentence_Number
+ (1 + p_kr | Participant_ID) + (1 + p_kr | Word_Unique_ID) + p_kr
   Data: eye_tracking (Number of observations: 2462)
Samples: 6 chains, each with iter = 3000; warmup = 1500; thin = 1;
         total post-warmup samples = 9000
    ICs: LOO = NA; WAIC = NA; R2 = NA

Group-Level Effects:
~Participant_ID (Number of levels: 84)
                    Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)           5.51      0.93     3.54     7.20       4809 1.00
sd(p_kr)                0.75      0.57     0.03     2.15       9000 1.00
cor(Intercept,p_kr)    -0.09      0.43    -0.82     0.74       9000 1.00

~Word_Unique_ID (Number of levels: 62)
                    Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)           2.57      1.18     0.26     4.69       2433 1.00
sd(p_kr)                1.14      0.81     0.05     2.99       5126 1.00
cor(Intercept,p_kr)     0.24      0.43    -0.68     0.89       9000 1.00

Population-Level Effects:
                       Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept                131.54      1.74   128.15   134.94       7551 1.00
Word_Length                0.89      0.76    -0.62     2.40       9000 1.00
Word_In_Sentence_Number   -1.03      0.75    -2.48     0.44       9000 1.00
p_kr                      -1.48      0.76    -2.97    -0.04       9000 1.00

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma    26.92      1.24    24.56    29.35       9000 1.00
beta     73.08      2.06    69.05    77.13       9000 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

B.1.3 P(referent|pronoun) Regression Path Duration

```
Family: exgaussian
  Links: mu = identity; sigma = identity; beta = identity
Formula: IA_REGRESSION_PATH_DURATION ~ Word_Length + Word_In_Sentence_Number +
(1 + p_kr | Participant_ID) + (1 + p_kr | Word_Unique_ID) + p_kr
   Data: eye_tracking (Number of observations: 2462)
Samples: 6 chains, each with iter = 3000; warmup = 1500; thin = 1;
         total post-warmup samples = 9000
    ICs: LOO = NA; WAIC = NA; R2 = NA

Group-Level Effects:
~Participant_ID (Number of levels: 84)
                    Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)           1.04      0.78     0.04     2.86       6943 1.00
sd(p_kr)                0.74      0.55     0.03     2.05       9000 1.00
cor(Intercept,p_kr)     0.00      0.46    -0.82     0.83       9000 1.00

~Word_Unique_ID (Number of levels: 62)
                    Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)           2.58      1.28     0.20     4.98       3548 1.00
sd(p_kr)                1.27      0.90     0.06     3.32       5738 1.00
cor(Intercept,p_kr)     0.22      0.43    -0.69     0.88       9000 1.00

Population-Level Effects:
                       Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept                118.50      1.67   115.21   121.78       9000 1.00
Word_Length                0.67      0.81    -0.92     2.28       9000 1.00
Word_In_Sentence_Number    0.11      0.83    -1.54     1.71       9000 1.00
p_kr                      -1.12      0.83    -2.77     0.52       9000 1.00

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma    21.79      1.30    19.27    24.37       9000 1.00
beta    166.67      3.64   159.72   173.85       9000 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

B.1.4 P(referent|pronoun) Total Number Fixations

```
Family: poisson
  Links: mu = log
Formula: IA_FIXATION_COUNT ~ Word_Length + Word_In_Sentence_Number +
(1 + p_kr | Participant_ID) + (1 + p_kr | Word_Unique_ID) + p_kr
   Data: eye_tracking (Number of observations: 2462)
Samples: 6 chains, each with iter = 3000; warmup = 1500; thin = 1;
         total post-warmup samples = 9000
    ICs: LOO = NA; WAIC = NA; R2 = NA

Group-Level Effects:
~Participant_ID (Number of levels: 84)
                    Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)           0.02      0.02     0.00     0.06       7005 1.00
sd(p_kr)                0.02      0.01     0.00     0.05       9000 1.00
cor(Intercept,p_kr)    -0.01      0.44    -0.82     0.81       9000 1.00

~Word_Unique_ID (Number of levels: 62)
                    Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)           0.03      0.02     0.00     0.08       4166 1.00
sd(p_kr)                0.04      0.03     0.00     0.10       4127 1.00
cor(Intercept,p_kr)    -0.02      0.44    -0.81     0.81       6752 1.00

Population-Level Effects:
                       Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept                  0.18      0.02     0.14     0.22       9000 1.00
Word_Length                0.02      0.02    -0.02     0.06       9000 1.00
Word_In_Sentence_Number    0.03      0.02    -0.01     0.07       9000 1.00
p_kr                      -0.01      0.02    -0.06     0.03       9000 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

B.2 Model's output for P(referent)

B.2.1 P(referent) Skipping Rate

```
 Family: bernoulli
  Links: mu = logit
Formula: IA_SKIP ~ Word_Length + Word_In_Sentence_Number
+ (1 + p_ref | Participant_ID) + (1 + p_ref | Word_Unique_ID) + p_ref
   Data: eye_tracking (Number of observations: 2402)
Samples: 6 chains, each with iter = 3000; warmup = 1500; thin = 1;
         total post-warmup samples = 9000
    ICs: LOO = NA; WAIC = NA; R2 = NA

Group-Level Effects:
~Participant_ID (Number of levels: 84)
                     Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)            0.78      0.10     0.60     0.99       3648 1.00
sd(p_ref)                0.13      0.09     0.01     0.33       2500 1.00
cor(Intercept,p_ref)    -0.18      0.38    -0.82     0.66       9000 1.00

~Word_Unique_ID (Number of levels: 60)
                     Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)            0.77      0.10     0.58     0.99       3542 1.00
sd(p_ref)                0.16      0.12     0.01     0.45       1950 1.00
cor(Intercept,p_ref)     0.12      0.46    -0.77     0.87       4457 1.00

Population-Level Effects:
                       Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept                 -1.57      0.15    -1.87    -1.28       3463 1.00
Word_Length               -0.28      0.13    -0.53    -0.03       3793 1.00
Word_In_Sentence_Number   -0.30      0.12    -0.54    -0.07       3442 1.00
p_ref                     -0.15      0.12    -0.39     0.09       4074 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

B.2.2 P(referent) First Fixation Duration

```
 Family: exgaussian
  Links: mu = identity; sigma = identity; beta = identity
Formula: IA_FIRST_FIXATION_DURATION ~ Word_Length + Word_In_Sentence_Number +
(1 + p_ref | Participant_ID) + (1 + p_ref | Word_Unique_ID) + p_ref
   Data: eye_tracking (Number of observations: 2402)
Samples: 6 chains, each with iter = 3000; warmup = 1500; thin = 1;
         total post-warmup samples = 9000
    ICs: LOO = NA; WAIC = NA; R2 = NA

Group-Level Effects:
~Participant_ID (Number of levels: 84)
                     Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)            5.18      1.00     2.95     6.95       3349 1.00
sd(p_ref)                0.72      0.54     0.03     1.99       9000 1.00
cor(Intercept,p_ref)    -0.10      0.44    -0.85     0.76       9000 1.00

~Word_Unique_ID (Number of levels: 60)
                     Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)            2.44      1.24     0.19     4.72       2750 1.00
sd(p_ref)                1.19      0.85     0.05     3.12       4206 1.00
cor(Intercept,p_ref)     0.10      0.44    -0.76     0.85       9000 1.00

Population-Level Effects:
                       Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept                131.19      1.73   127.80   134.60       7740 1.00
Word_Length                0.96      0.77    -0.56     2.47       9000 1.00
Word_In_Sentence_Number   -1.07      0.76    -2.56     0.43       9000 1.00
p_ref                     -0.50      0.78    -2.02     1.05       9000 1.00

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma    27.24      1.25    24.87    29.74       9000 1.00
beta     72.93      2.07    68.96    77.09       9000 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

B.2.3 P(referent) Regression Path Duration

```
Family: exgaussian
  Links: mu = identity; sigma = identity; beta = identity
Formula: IA_REGRESSION_PATH_DURATION ~ Word_Length + Word_In_Sentence_Number
+ (1 + p_ref | Participant_ID) + (1 + p_ref | Word_Unique_ID) + p_ref
   Data: eye_tracking (Number of observations: 2402)
Samples: 6 chains, each with iter = 3000; warmup = 1500; thin = 1;
         total post-warmup samples = 9000
    ICs: LOO = NA; WAIC = NA; R2 = NA

Group-Level Effects:
~Participant_ID (Number of levels: 84)
                     Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)            1.03      0.77     0.04     2.84       5499 1.00
sd(p_ref)                0.74      0.56     0.03     2.10       9000 1.00
cor(Intercept,p_ref)    -0.02      0.45    -0.83     0.80       9000 1.00

~Word_Unique_ID (Number of levels: 60)
                     Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)            2.27      1.33     0.12     4.88       2622 1.00
sd(p_ref)                1.71      1.12     0.08     4.09       3199 1.00
cor(Intercept,p_ref)    -0.02      0.42    -0.79     0.78       7128 1.00

Population-Level Effects:
                       Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept                118.21      1.69   114.90   121.52       9000 1.00
Word_Length                0.72      0.81    -0.89     2.30       9000 1.00
Word_In_Sentence_Number    0.05      0.82    -1.55     1.66       9000 1.00
p_ref                     -0.32      0.84    -1.98     1.32       9000 1.00

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma    22.01      1.32    19.48    24.67       9000 1.00
beta    166.98      3.72   159.95   174.57       9000 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

B.2.4 P(referent) Total Number Fixations

```
 Family: poisson
  Links: mu = log
Formula: IA_FIXATION_COUNT ~ Word_Length + Word_In_Sentence_Number +
(1 + p_ref | Participant_ID) + (1 + p_ref | Word_Unique_ID) + p_ref
   Data: eye_tracking (Number of observations: 2402)
Samples: 6 chains, each with iter = 3000; warmup = 1500; thin = 1;
         total post-warmup samples = 9000
    ICs: LOO = NA; WAIC = NA; R2 = NA

Group-Level Effects:
~Participant_ID (Number of levels: 84)
                     Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)            0.02      0.02     0.00     0.06       9000 1.00
sd(p_ref)                0.02      0.01     0.00     0.05       9000 1.00
cor(Intercept,p_ref)    -0.02      0.44    -0.81     0.79       9000 1.00

~Word_Unique_ID (Number of levels: 60)
                     Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)            0.03      0.02     0.00     0.08       5374 1.00
sd(p_ref)                0.03      0.02     0.00     0.09       4186 1.00
cor(Intercept,p_ref)     0.01      0.44    -0.80     0.79       9000 1.00

Population-Level Effects:
                       Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept                  0.19      0.02     0.15     0.23       9000 1.00
Word_Length                0.02      0.02    -0.02     0.06       9000 1.00
Word_In_Sentence_Number    0.03      0.02    -0.01     0.07       9000 1.00
p_ref                     -0.02      0.02    -0.06     0.02       9000 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

B.3 Model's output for P(answer)

B.3.1 P(answer) Skipping Rate

```
 Family: bernoulli
  Links: mu = logit
Formula: IA_SKIP ~ Word_Length + Word_In_Sentence_Number + (1 + p_ans | Participant_ID)
+ (1 + p_ans | Word_Unique_ID) + p_ans
   Data: eye_tracking (Number of observations: 2646)
Samples: 6 chains, each with iter = 3000; warmup = 1500; thin = 1;
         total post-warmup samples = 9000
    ICs: LOO = NA; WAIC = NA; R2 = NA

Group-Level Effects:
~Participant_ID (Number of levels: 84)
                     Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)            0.73      0.09     0.56     0.92       3562 1.00
sd(p_ans)                0.13      0.09     0.01     0.33       2799 1.00
cor(Intercept,p_ans)     0.23      0.37    -0.60     0.84       9000 1.00

~Word_Unique_ID (Number of levels: 65)
                     Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)            0.81      0.11     0.61     1.03       2756 1.00
sd(p_ans)                0.23      0.17     0.01     0.61       1012 1.00
cor(Intercept,p_ans)    -0.04      0.39    -0.77     0.73       4974 1.00

Population-Level Effects:
                       Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept                 -1.49      0.15    -1.78    -1.21       3001 1.00
Word_Length               -0.28      0.12    -0.52    -0.04       4125 1.00
Word_In_Sentence_Number   -0.26      0.12    -0.50    -0.01       3368 1.00
p_ans                     -0.27      0.13    -0.53    -0.01       3867 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

B.3.2 P(answer) First Fixation Duration

```
 Family: exgaussian
  Links: mu = identity; sigma = identity; beta = identity
Formula: IA_FIRST_FIXATION_DURATION ~ Word_Length + Word_In_Sentence_Number +
(1 + p_ans | Participant_ID) + (1 + p_ans | Word_Unique_ID) + p_ans
   Data: eye_tracking (Number of observations: 2646)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000
    ICs: LOO = NA; WAIC = NA; R2 = NA

Group-Level Effects:
~Participant_ID (Number of levels: 84)
                     Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)            5.71      0.85     4.00     7.28       3014 1.00
sd(p_ans)                0.69      0.54     0.02     2.02       4000 1.00
cor(Intercept,p_ans)    -0.04      0.44    -0.81     0.79       4000 1.00

~Word_Unique_ID (Number of levels: 65)
                     Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)            2.19      1.22     0.15     4.52       1447 1.00
sd(p_ans)                1.53      0.88     0.07     3.28       1989 1.00
cor(Intercept,p_ans)     0.19      0.42    -0.69     0.87       4000 1.00

Population-Level Effects:
                       Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept                131.76      1.67   128.56   135.06       4000 1.00
Word_Length                0.68      0.69    -0.65     2.00       4000 1.00
Word_In_Sentence_Number   -0.75      0.77    -2.24     0.73       4000 1.00
p_ans                     -1.08      0.72    -2.50     0.34       4000 1.00

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma    26.87      1.14    24.72    29.08       4000 1.00
beta     72.97      1.99    69.14    76.85       4000 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

B.3.3 P(answer) Regression Path Duration

```
 Family: exgaussian
  Links: mu = identity; sigma = identity; beta = identity
Formula: IA_REGRESSION_PATH_DURATION ~ Word_Length + Word_In_Sentence_Number + (1 + p_ans
| Participant_ID) + (1 + p_ans | Word_Unique_ID) + p_ans
   Data: eye_tracking (Number of observations: 2646)
Samples: 6 chains, each with iter = 3000; warmup = 1500; thin = 1;
         total post-warmup samples = 9000
    ICs: LOO = NA; WAIC = NA; R2 = NA

Group-Level Effects:
~Participant_ID (Number of levels: 84)
                     Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)            1.09      0.81     0.04     3.01       5664 1.00
sd(p_ans)                0.77      0.59     0.03     2.19       9000 1.00
cor(Intercept,p_ans)    -0.01      0.45    -0.82     0.82       9000 1.00

~Word_Unique_ID (Number of levels: 65)
                     Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)            2.82      1.34     0.22     5.18       2639 1.00
sd(p_ans)                1.32      0.89     0.06     3.28       4701 1.00
cor(Intercept,p_ans)     0.11      0.43    -0.73     0.85       9000 1.00

Population-Level Effects:
                       Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept                118.30      1.61   115.15   121.47       9000 1.00
Word_Length                1.37      0.78    -0.18     2.92       9000 1.00
Word_In_Sentence_Number   -0.18      0.81    -1.76     1.43       9000 1.00
p_ans                     -0.29      0.81    -1.87     1.27       9000 1.00

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma    21.55      1.24    19.18    24.09       9000 1.00
beta    170.89      3.63   163.99   178.13       9000 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

B.3.4 P(answer) Total Number Fixations

```
Family: poisson
  Links: mu = log
Formula: IA_FIXATION_COUNT ~ Word_Length + Word_In_Sentence_Number +
(1 + p_ans | Participant_ID) + (1 + p_ans | Word_Unique_ID) + p_ans
   Data: eye_tracking (Number of observations: 2646)
Samples: 6 chains, each with iter = 3000; warmup = 1500; thin = 1;
         total post-warmup samples = 9000
    ICs: LOO = NA; WAIC = NA; R2 = NA

Group-Level Effects:
~Participant_ID (Number of levels: 84)
                     Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)            0.02      0.02     0.00     0.06       7209 1.00
sd(p_ans)                0.02      0.01     0.00     0.04       9000 1.00
cor(Intercept,p_ans)    -0.01      0.45    -0.82     0.81       9000 1.00

~Word_Unique_ID (Number of levels: 65)
                     Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)            0.03      0.02     0.00     0.08       4362 1.00
sd(p_ans)                0.05      0.02     0.00     0.10       2836 1.00
cor(Intercept,p_ans)     0.01      0.44    -0.79     0.81       4681 1.00

Population-Level Effects:
                       Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept                  0.19      0.02     0.15     0.23       9000 1.00
Word_Length                0.05      0.02     0.02     0.09       9000 1.00
Word_In_Sentence_Number    0.01      0.02    -0.03     0.05       9000 1.00
p_ans                      0.01      0.02    -0.03     0.06       9000 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Bibliography

Abeillé, Anne, Lionel Clément, and François Toussenel (2003). "Building a treebank for French". In: *Treebanks*. Springer, pp. 165–187.

Agic, Željko et al. (2015). "Universal dependencies 1.1". In: *LINDAT/CLARIN Digital Library at Institute of Formal and Applied Linguistics, Charles University in Prague* 3.

Ariel, Mira (1988). "Referring and accessibility". In: *Journal of Linguistics* 24.01, pp. 65–87.

— (1991). "The function of accessibility in a theory of grammar". In: *Journal of Pragmatics* 16.5, pp. 443–463.

Artstein, Ron and Massimo Poesio (2008). "Inter-coder agreement for computational linguistics". In: *Computational Linguistics* 34.4, pp. 555–596.

Barrett, Maria, Željko Agić, and Anders Søgaard (2015). "The Dundee Treebank". In: *Proceedings of the 14th International Workshop on Treebanks and Linguistic Theories (TLT 14)*.

Barzilay, Regina and Mirella Lapata (2008). "Modeling local coherence: An entity-based approach". In: *Computational Linguistics* 34.1, pp. 1–34.

Bates, Douglas et al. (2015). "Fitting linear mixed-effects models using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48. DOI: 10.18637/jss.v067.i01.

Bayes, Mr. and Mr Price (1763). "An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS". In: *Philosophical Transactions (1683-1775)*, pp. 370–418.

Beniamine, Sacha (2018). "Classifications flexionnelles : Étude quantitative des structures de paradigmes". PhD thesis. Université Sorbonne Paris Cité-Université Paris Diderot (Paris 7).

Bergsma, Shane and Dekang Lin (2006). "Bootstrapping path-based pronoun resolution". In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 33–40.

Bird, Steven, Ewan Klein, and Edward Loper (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. " O'Reilly Media, Inc."

Bonferroni, Carlo E (1936). "Teoria statistica delle classi e calcolo delle probabilita". In: *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* 8. Ed. by Libreria internazionale Seeber.

Bott, Oliver and Torgrim Solstad (2014). "From verbs to discourse: A novel account of implicit causality". In: *Psycholinguistic Approaches to Meaning and Understanding Across Languages*. Springer, pp. 213–251.

Brennan, Jonathan R et al. (2016). "Abstract linguistic structure correlates with temporal activity during naturalistic comprehension". In: *Brain and Language* 157, pp. 81–94.

Brennan, Susan E, Marilyn W Friedman, and Carl J Pollard (1987). "A centering approach to pronouns". In: *Proceedings of the 25th Annual Meeting of the Association*

for Computational Linguistics. Association for Computational Linguistics, pp. 155–162.

Broadbent, Donald E (1970). "In defense of empirical psychology". In: *Bulletin of the British Psychological Society*.

Bürkner, Paul-Christian (2017). "brms: An R Package for Bayesian Multilevel Models Using Stan". In: *Journal of Statistical Software* 80.1, pp. 1–28. DOI: 10.18637/jss.v080.i01.

Clancy, Patricia M (1980). "Referential choice in English and Japanese narrative discourse". In: *The pear stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production* 3, pp. 127–201.

Clark, Herbert H and CJ Sengul (1979). "In search of referents for nouns and pronouns". In: *Memory & Cognition* 7.1, pp. 35–41.

Clark, Kevin and Christopher D Manning (2016a). "Improving Coreference Resolution by Learning Entity-Level Distributed Representations". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1, pp. 643–653.

— (2016b). "Deep reinforcement learning for mention-ranking coreference models". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2256–2262.

Cohen, Jacob (1960). "A coefficient of agreement for nominal scales". In: *Educational and Psychological Measurement* 20.1, pp. 37–46.

Cop, Uschi et al. (2017). "Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading". In: *Behavior Research Methods* 49.2, pp. 602–615.

Crawley, Rosalind A, Rosemary J Stevenson, and David Kleinman (1990). "The use of heuristic strategies in the interpretation of pronouns". In: *Journal of Psycholinguistic Research* 19.4, pp. 245–264.

Demberg, Vera and Frank Keller (2008). "Data from eye-tracking corpora as evidence for theories of syntactic processing complexity". In: *Cognition* 109.2, pp. 193–210.

Denis, Pascal and Jason Baldridge (2007). "A ranking approach to pronoun resolution". In: *Proceedings of the International Joint Conferences on Artificial Intelligence*, pp. 1588–1593.

Downey, Allen (2013). *Think Bayes: Bayesian Statistics in Python*. " O'Reilly Media, Inc."

Dubey, Amit, Frank Keller, and Patrick Sturt (2013). "Probabilistic modeling of discourse-aware sentence processing". In: *Topics in Cognitive Science* 5.3, pp. 425–451.

Eager, Christopher and Joseph Roy (2017). "Mixed effects models are sometimes terrible". In: *arXiv preprint arXiv:1701.04858*.

Ehrlich, Kate and Keith Rayner (1983). "Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing". In: *Journal of Verbal Learning and Verbal Behavior* 22.1, pp. 75–87.

Elsner, Micha and Eugene Charniak (2008). "Coreference-inspired coherence modeling". In: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Association for Computational Linguistics, pp. 41–44.

— (2011). "Disentangling chat with local coherence models". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pp. 1179–1189.

Field, Andy (2009). *Discovering Statistics Using SPSS*. Sage publications.

Frank, Stefan L (2013). "Uncertainty reduction as a measure of cognitive load in sentence comprehension". In: *Topics in Cognitive Science* 5.3, pp. 475–494.

Frank, Stefan L and Rens Bod (2011). "Insensitivity of the human sentence-processing system to hierarchical structure". In: *Psychological Science* 22.6, pp. 829–834.

Frank, Stefan L et al. (2007). "Coherence-driven resolution of referential ambiguity: A computational model". In: *Memory & Cognition* 35.6, pp. 1307–1322.

Frank, Stefan L et al. (2009). "Surprisal-based comparison between a symbolic and a connectionist model of sentence processing". In: *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Cognitive Science Society Austin, TX, pp. 1139–1144.

Frederiksen, John R (1981). "Understanding anaphora: Rules used by readers in assigning pronominal referents". In: *Discourse Processes* 4.4, pp. 323–347.

Fukumura, Kumiko and Roger PG van Gompel (2015). "Effects of order of mention and grammatical role on anaphor resolution". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 41.2, p. 501.

Futrell, Richard et al. (2018). "The Natural Stories Corpus". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). ISBN: 979-10-95546-00-9.

Gelman, Andrew, Jennifer Hill, and Masanao Yajima (2012). "Why we (usually) don't have to worry about multiple comparisons". In: *Journal of Research on Educational Effectiveness* 5.2, pp. 189–211.

Gernsbacher, Morton Ann (1990). *Language Comprehension as Structure Building*.

Gernsbacher, Morton Ann and David J Hargreaves (1988). "Accessing sentence participants: The advantage of first mention". In: *Journal of Memory and Language* 27.6, pp. 699–717.

Gernsbacher, Morton Ann, David J Hargreaves, and Mark Beeman (1989). "Building and accessing clausal representations: The advantage of first mention versus the advantage of clause recency". In: *Journal of Memory and Language* 28.6, pp. 735–755.

Gibson, Edward (2000). "The dependency locality theory: A distance-based theory of linguistic complexity". In: *Image, Language, Brain*, pp. 95–126.

Gordon, Peter C, Barbara J Grosz, and Laura A Gilliom (1993). "Pronouns, names, and the centering of attention in discourse". In: *Cognitive Science* 17.3, pp. 311–347.

Grenander, Ulf (1967). *Syntax-Controlled Probabilities*. Division of Applied Mathematics, Brown University.

Grosz, Barbara J, Aravind K Joshi, and Scott Weinstein (1983). "Providing a unified account of definite noun phrases in discourse". In: *Proceedings of the 21st Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 44–50.

Grosz, Barbara J, Scott Weinstein, and Aravind K Joshi (1995). "Centering: A framework for modeling the local coherence of discourse". In: *Computational Linguistics* 21.2, pp. 203–225.

Hale, John (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*. Association for Computational Linguistics, pp. 1–8.

Hale, John (2003). "The information conveyed by words in sentences". In: *Journal of Psycholinguistic Research* 32.2, pp. 101–123.

— (2006). "Uncertainty about the rest of the sentence". In: *Cognitive Science* 30.4, pp. 643–672.

Hemforth, Barbara et al. (2010). "Language specific preferences in anaphor resolution: Exposure or gricean maxims?" In: *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pp. 2218–2223.

Hobbs, Jerry R (1976). *Pronoun Resolution (Research Report 76-1)*.

— (1978). "Resolving pronoun references". In: *Lingua* 44.4, pp. 311–338.

— (1979). "Coherence and coreference". In: *Cognitive Science* 3.1, pp. 67–90.

Jaeger, T Florian (2010). "Redundancy and reduction: Speakers manage syntactic information density". In: *Cognitive Psychology* 61.1, pp. 23–62.

Jaffe, Evan, Cory Shain, and William Schuler (2018). "Coreference and focus in reading times". In: *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pp. 1–9.

Järvikivi, Juhani et al. (2005). "Ambiguous pronoun resolution contrasting the first-mention and subject-preference accounts". In: *Psychological Science* 16.4, pp. 260–264.

Ji, Yangfeng et al. (2017). "Dynamic entity representations in neural language models". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1830–1839.

Kaiser, Elsi and John C Trueswell (2008). "Interpreting pronouns and demonstratives in Finnish: Evidence for a form-specific approach to reference resolution". In: *Language and Cognitive Processes* 23.5, pp. 709–748.

Kehler, Andrew and Hannah Rohde (2013). "A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation". In: *Theoretical Linguistics* 39.1-2, pp. 1–37.

— (2018). "Prominence and coherence in a Bayesian theory of pronoun interpretation". In: *Journal of Pragmatics*.

Kennedy, Alan, Robin Hill, and Joël Pynte (2003). "The Dundee Corpus". In: *Proceedings of the 12th European Conference on Eye Movement*.

Kennedy, Alan and Joël Pynte (2005). "Parafoveal-on-foveal effects in normal reading". In: *Vision Research* 45.2, pp. 153–168.

Klein, Dan and Christopher D. Manning (2003). "Accurate unlexicalized parsing". In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, pp. 423–430.

Kravtchenko, Ekaterina (2014). "Predictability and syntactic production: Evidence from subject omission in Russian". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 36. 36.

Krippendorff, Klaus (1980). *Content Analysis: An Introduction to its Methodology*. Sage Publications.

Lappin, Shalom and Herbert J Leass (1994). "An algorithm for pronominal anaphora resolution". In: *Computational Linguistics* 20.4, pp. 535–561.

Lee, Heeyoung et al. (2011). "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task". In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*. Association for Computational Linguistics, pp. 28–34.

Lee, Kenton et al. (2017). "End-to-end neural coreference resolution". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 188–197.

Lewandowski, Daniel, Dorota Kurowicka, and Harry Joe (2009). "Generating random correlation matrices based on vines and extended onion method". In: *Journal of Multivariate Analysis* 100.9, pp. 1989–2001.

Linzen, Tal and T Florian Jaeger (2014). "Investigating the role of entropy in sentence processing". In: *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics*, pp. 10–18.

Luke, Steven G and Kiel Christianson (2016). "Limits on lexical prediction during reading". In: *Cognitive Psychology* 88, pp. 22–60.

Luo, Xiaoqiang and Sameer Pradhan (2016). "Evaluation metrics". In: *Anaphora Resolution*. Ed. by Massimo Poesio, Roland Stuckardt, and Yannick Versley. Springer, pp. 141–163.

Luo, Xiaoqiang et al. (2004). "A mention-synchronous coreference resolution algorithm based on the bell tree". In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 135–143.

Manning, Christopher et al. (2014). "The Stanford CoreNLP natural language processing toolkit". In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60.

Maratsos, Michael P (1973). "The effects of stress on the understanding of pronominal co-reference in children". In: *Journal of Psycholinguistic Research* 2.1, pp. 1–8.

Marcus, Mitchell P, Mary Ann Marcinkiewicz, and Beatrice Santorini (1993). "Building a large annotated corpus of English: The Penn Treebank". In: *Computational Linguistics* 19.2, pp. 313–330.

Martin, James H and Daniel Jurafsky (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson/Prentice Hall.

Matuschek, Hannes et al. (2017). "Balancing type I error and power in linear mixed models". In: *Journal of Memory and Language* 94, pp. 305–315.

McRae, Ken, Michael J Spivey-Knowlton, and Michael K Tanenhaus (1998). "Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension". In: *Journal of Memory and Language* 38.3, pp. 283–312.

Mitchell, Jeff et al. (2010). "Syntactic and semantic factors in processing difficulty: An integrated measure". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 196–206.

Mitkov, Ruslan (1998). "Robust pronoun resolution with limited knowledge". In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. Vol. 2. Association for Computational Linguistics, pp. 869–875.

— (2002). *Anaphora Resolution*. Longman.

Modi, Ashutosh et al. (2016). "InScript: Narrative texts annotated with script information". In: *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*.

Modi, Ashutosh et al. (2017). "Modelling semantic expectation: using script knowledge for referent prediction". In: *Transactions of the Association of Computational Linguistics* 5.1, pp. 31–44.

Muthén, Bengt and Tihomir Asparouhov (2012). "Bayesian structural equation modeling: a more flexible representation of substantive theory". In: *Psychological Methods* 17.3, p. 313.

Muzerelle, Judith et al. (2014). "ANCOR_Centre, a large free spoken French coreference corpus: description of the resource and reliability measures". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 843–847.

Ng, Vincent and Claire Cardie (2002). "Improving machine learning approaches to coreference resolution". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 104–111.

Oberle, Bruno (2018). "SACR: A drag-and-drop based tool for coreference annotation". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). ISBN: 979-10-95546-00-9.

Passonneau, Rebecca (2006). "Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation". In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation*.

Passonneau, Rebecca J (2004). "Computing reliability for coreference annotation." In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation*.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Piantadosi, Steven T, Harry Tily, and Edward Gibson (2011). "Word lengths are optimized for efficient communication". In: *Proceedings of the National Academy of Sciences* 108.9, pp. 3526–3529.

Poesio, Massimo, Roland Stuckardt, and Yannick Versley, eds. (2016). *Anaphora Resolution. Algorithms, resources, and applications*. Theory and applications of natural language processing. Springer. ISBN: 9783662479087.

Poesio, Massimo et al. (2016). "Annotated corpora and annotation tools". In: *Anaphora Resolution*. Ed. by Massimo Poesio, Roland Stuckardt, and Yannick Versley. Springer, pp. 97–140.

Pradhan, Sameer et al. (2011). "Conll-2011 shared task: Modeling unrestricted coreference in ontonotes". In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*. Association for Computational Linguistics, pp. 1–27.

Pynte, Joël and Saveria Colonna (2000). "Decoupling syntactic parsing from visual inspection: The case of relative clause attachment in French". In: *Reading as a Perceptual Process*, pp. 529–547.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org.

Rayner, Keith (1998). "Eye movements in reading and information processing: 20 years of research". In: *Psychological Bulletin* 124.3, p. 372.

Recasens, Marta and Eduard Hovy (2009). "A deeper look into features for coreference resolution". In: *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium*. Springer, pp. 29–42.

Rondal, Jean A et al. (1984). "Coréférence et stratégie des fonctions parallèles dans le cas des pronoms anaphoriques ambigus." In: *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*.

Seminck, Olga (2016). "Un modèle simple du coût cognitif de la résolution des anaphores". In: *Actes de la 18ième Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pp. 66–79.

Seminck, Olga and Pascal Amsili (2017). "A computational model of human preferences for pronoun resolution". In: *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 53–63.

— (2018). "A Gold Anaphora Annotation Layer on an Eye Movement Corpus". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari (Conference chair) et al. Miyazaki, Japan: European Language Resources Association (ELRA). ISBN: 979-10-95546-00-9.

Shannon, Claude E (1948). "A mathematical theory of communication". In: *Bell System Technical Journal* 27, pp. 379–423.

Shannon, Claude E and Warren Weaver (1949). *The mathematical theory of communication*.

Sheldon, Amy (1974). "The role of parallel function in the acquisition of relative clauses in English". In: *Journal of Verbal Learning and Verbal Behavior* 13.3, pp. 272–281.

Sheldon, Ross (1998). *A First Course in Probability*. Prentice-Hall.

Shillcock, Richard (1982). "The on-line resolution of pronominal anaphora". In: *Language and Speech* 25.4, pp. 385–401.

Smyth, Ron (1994). "Grammatical determinants of ambiguous pronoun resolution". In: *Journal of Psycholinguistic Research* 23.3, pp. 197–229.

Soon, Wee Meng, Hwee Tou Ng, and Daniel Chung Yong Lim (2001). "A machine learning approach to coreference resolution of noun phrases". In: *Computational Linguistics* 27.4, pp. 521–544.

Staub, Adrian (2011). "The effect of lexical predictability on distributions of eye fixation durations". In: *Psychonomic Bulletin & Review* 18.2, pp. 371–376.

Stevenson, Rosemary J, Rosalind A Crawley, and David Kleinman (1994). "Thematic roles, focus and the representation of events". In: *Language and Cognitive Processes* 9.4, pp. 519–548.

Stolcke, Andreas (1995). "An efficient probabilistic context-free parsing algorithm that computes prefix probabilities". In: *Computational Linguistics* 21.2, pp. 165–201.

Stuckardt, Roland (2016). "Introduction. Algorithms, resources, and applications". In: *Anaphora Resolution*. Ed. by Massimo Poesio, Roland Stuckardt, and Yannick Versley. Theory and applications of natural language processing. Literaturangaben. Springer, pp. 1–19. ISBN: 9783662479087.

Tabor, Whitney and Michael K Tanenhaus (1999). "Dynamical models of sentence processing". In: *Cognitive Science* 23.4, pp. 491–515.

Tetreault, Joel R (2001). "A corpus-based evaluation of centering and pronoun resolution". In: *Computational Linguistics* 27.4, pp. 507–520.

Thomas, Joy A and TM Cover (2006). *Elements of Information Theory*. 2nd ed. Wiley New York.

Tily, Harry and Steven Piantadosi (2009). "Refer efficiently: Use less informative expressions for more predictable meanings". In: *Proceedings of the Workshop on the Production of Referring Expressions: Bridging the Gap Between Computational and Empirical Approaches to Reference*.

Van Deemter, Kees and Rodger Kibble (2000). "On coreferring: Coreference in MUC and related annotation schemes". In: *Computational Linguistics* 26.4, pp. 629–637.

Van Gompel, Roger PG and Asifa Majid (2004). "Antecedent frequency effects during the processing of pronouns". In: *Cognition* 90.3, pp. 255–264.

Vilain, Marc et al. (1995). "A model-theoretic coreference scoring scheme". In: *Proceedings of the 6th Conference on Message Understanding*. Association for Computational Linguistics, pp. 45–52.

Von der Malsburg, Titus (2018). "The president will give her inauguration speech: Explicit belief and implicit expectations in language production and comprehension". In: *Forum Entwicklung und Anwendung von Sprach-Technologien (Oral Presentation)*. Universität des Saarlandes.

Von der Malsburg, Titus and B Angele (2017). "False positives and other statistical errors in standard analyses of eye movements in reading". In: *Journal of Memory and Language* 94, p. 119.

Wiseman, Sam, Alexander M Rush, and Stuart M Shieber (2016). "Learning global features for coreference resolution". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 994–1004.

Wiseman, Sam et al. (2015). "Learning anaphoricity and antecedent ranking features for coreference resolution". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Vol. 1, pp. 1416–1426.

Yang, Xiaofeng et al. (2004). "Improving pronoun resolution by incorporating coreferential information of candidates". In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, p. 127.