Formal approaches to the communication of conflictual social identities in discourse

Quentin Dénigot

Université de Paris
ED 622 - Sciences du langage
Laboratoire de Linguistique Formelle

FORMAL APPROACHES TO THE COMMUNICATION OF CONFLICTING IDENTITIES IN DISCOURSE

QUENTIN DÉNIGOT

Thèse de doctorat de Linguistique

Dirigée par:
Heather BURNETT, PHD, HDR, DR CNRS
Présentée et soutenue publiquement le 11 Avril 2022

Devant un jury composé de:
Jonathan GINZBURG (PU), Université de Paris, Président
Elin MCCREADY (PROFESSOR), Aoyama Gakuin University, Rapporteure
Jennifer SAUL (PROFESSOR), University of Waterloo, Rapporteure
Denis PAPERNO (Assistant Professor), Utrecht University, Examinateur
Nicholas ASHER (DR CNRS), Université de Toulouse 3, Examinateur

Quentin Dénigot: *Formal Approaches to the Communication of Conflicting Identities in Discourse, ,* 11 Avril 2022

Ce travail est dédié à toustes les professeur·e·s qui m'ont permis d'en arriver là. Ce travail est aussi le fruit du leur.

ABSTRACT

FORMAL APPROACHES TO THE COMMUNICATION OF CONFLICTUAL SOCIAL IDENTITIES IN DISCOURSE

This dissertation focuses on a proposal for the formal modelling of a very specific phenomenon of political communication usually called "dogwhistles". A secondary contribution is an attempt to use (and a reflexive approach to the use of) machine learning techniques for the study of said dogwhistles.

From a theoretical point of view, this work draws inspiration from game theoretic approaches of linguistics and pragmatics, game theoretic approaches of social meaning, and distributional semantics. Beyond the main contributions, the thesis presents briefly each of these areas of research, since it is likely that readers might not know much about at least one of these areas, and I therefore hope that these introductory texts can be also counted among the contributions of this work.

Formal linguistics has devised a number of sophisticated mathematical objects to try and describe linguistic phenomena with as much precision as possible for the last decades. There is however one area of linguistics that has historically evaded attempts at formalization: sociolinguistics. The few attempts at formalizing sociolinguistics results are fairly recent. Inspired by tools from formal pragmatics, they have expanded the concept of linguistic interpretation to the interpretation not only of message content, but of social meaning as well, leading to elaborate formal models that are at the basis of fascinating research programmes.

There has been however few attempts at linking the works from formal pragmatics and semantics – focusing mostly on the interpretation of content – and formal sociolinguistics – focusing on the interpretation of social meaning. Notably, there have been no attempts at giving a picture in formal terms of situations whereby the social meaning contained within an utterance can influence the content interpretation of that utterance by listeners.

In this work, we focus on situations where speakers send messages where the social meaning and content interpretations vary according to who receives them. This phenomenon is sometimes referred to as dogwhistle politics, since it has been mostly studied in the context of political discourse. Using formal tools from pragmatics (Rational Speech Act models) and their adaptations in the realm of sociolinguistics (Social Meaning Games), we aim to construct a model that can accurately describes these situations, where the retrieval of the identity of the speaker can supposedly lead to a variety of interpretations, linking the notions of social meaning and content meaning in a single kind of linguistic game. Beyond this endeavor, a key characteristic of dogwhistles is that they are supposedly hidden messages to a point. This

leads to a situation where it can be difficult to assess whether a given utterance does or does not contain a dogwhistle. Acknowledging this, we have tried to give a definition of dogwhistles that goes beyond their mere description and could be applied to the enterprise of finding them in text, using concepts from distributional semantics, vector space models of language and machine learning.

Keywords: *dogwhistle, game theory, pragmatics, social meaning, distributional semantics, discourse*

RÉSUMÉ

APPROCHES FORMELLES DE LA COMMUNICATION D'IDENTITÉS CONFLICTUELLES DANS LE DISCOURS

Cette thèse a pour objet la modélisation formelle d'un phénomène très spécifique de la communication politique généralement appelé "dogwhistle". Une seconde contribution de ce travail consiste en une tentative d'utilisation (ainsi qu'un réflexion sur cette utilisation) de techniques d'apprentissage automatique pour l'étude des dogwhistles en question.

D'un point de vue théorique, ce travail s'inspire d'approches en théorie des jeux en linguistique, notamment celles portant sur la notion de sens social, et de sémantique distributionnelle. Au-delà des contributions principales, la thèse présente brièvement chacun de ces domaines de recherche, puisqu'il est probable que les lecteurices manquent de familiarité avec au moins l'une de ces notions. J'espère donc que ces textes introductifs puissent être comptés parmi les contributions de la thèse.

La linguistique formelle a au cours des dernières décennies conçu un nombre important d'objets mathématiques sophistiqués pour tenter de décrire différents phénomènes linguistiques avec autant de précision que possible. Il existe cependant un champ de la linguistique ayant historiquement échappé aux tentatives de formalisation: la sociolinguistique. Les quelques tentatives de formalisation de résultats en sociolinguistique sont relativement récents. A l'aide d'outils issus de la pragmatique formelle, ces formalisation ont étendu la notion d'interprétation des messages linguistiques à leur interprétation non seulement en termes propositionnels, mais aussi en termes de sens social, menant à des modèles complexes qui sont à la base de programmes de recherche entiers.

Il n'y a cependant eu que peu de tentatives de relier les travaux en en pragmatique et sémantique formelles à ceux de la sociolinguistique formelle. En particulier, il n'y a pas eu de travaux donnant à voir une description en termes formels de situations dans lesquelles le sens social contenu dans un énoncé soit en mesure d'influencer son sens propositionnel dans la réception par les interlocuteurices.

Dans ce travail, nous nous concentrons sur des situations dans lesquelles des locuteurices envoient des messages dans lesquels le sens social et le sens propositionnel varient selon l'identité des interlocuteurices. Ce phénomène, qu'on pourrait appeler "politique du sifflet à chien" (dogwhistle politics) a été essentiellement étudié dans le cadre de la communication politique. En utilisant les outils formels de la pragmatique (en particulier les modèles basés sur le Rational Speech Act) et leurs adaptations dans le monde de la sociolinguistique (Social Meaning Games), nous tenterons de construire un modèle qui puisse adéquatement décrire ces situations dans lesquelles l'attribution d'une identité particulière à l'émetteurice d'un message conduit supposément à un ensemble d'interprétations différentes, liant par là les notions de sens social et de sens propositionnel en un unique jeu linguistique. Au-delà de cette tâche, une caractéristique clé des dogwhistle se trouve dans leur côté caché. Cela implique qu'il peut être difficile de savoir si un énoncé donné contient ou non un dogwhistle. Au vu de cette difficulté, nous avons également tenté de donner une définition des dogwhistles allant au-delà de leur simple description et pouvant être appliquée à l'entreprise de leur découverte dans un texte, grâce à l'utilisation de concepts issus de la sémantique dsitributionelle, des modèles vectoriels de langage et d'apprentissage automatique.

Mots-clés: *dogwhistle, théorie des jeux, pragmatique, sens social, sémantique distributionnelle, discours*

RÉSUMÉ LONG

"On ne sort de l'ambiguïté qu'à son détriment"

— Citation (probablement apocryphe) généralement attribuée aux mémoires du Cardinal de Retz. ¬(Retz, 1825/1717) ?

L'auteur de la citation précédente, quel qu'il fût, l'avait bien compris : en politique, ce que l'on gagne en clarté se perd en assentiment. Dans le cadre du discours politique en démocratie représentative, où il est primordial pour tout·e politicien·ne de s'assurer le soutien d'une majorité de la population, une stratégie de communication reposant sur l'ambiguïté, la plus à même de minimiser le mécontentement, devrait s'imposer d'elle-même. L'objet de cette thèse est d'aborder une telle stratégie de communication politique. Plus pécisément, l'objet de cette thèse est de présenter, décrire et expliquer une déclinaison particulière de ce genre de stratégie discursive appelée *dogwhistle politics*, 'politique du sifflet à chien', une pratique qui repose sur l'utilisation de termes, phrases, représentation iconographiques, etc... qui vont être interprétées d'une façon spécifique dans un sous-ensemble spécifique de la population (les 'chiens') et être interprétées d'une autre façon par le reste de la population.

Ainsi, la phrase suivante pourra entraîner des interprétations différentes en fonction de qui l'entend:

(1) Non, l'ennemi n'est pas le musulman, c'est le financier !

Pour la majorité de l'audience, cette phrase[1] ne sera pas spécialement remarquable au-delà d'être représentative d'un certain discours de gauche. Pour certains, "le financier" n'est autre qu'Emmanuel Macron, qui avant sa carrière politique s'est justement illustré dans le monde de la finance. Pour d'autres, enfin, la mise en avant de la figure d'une personne de foi ("le musulman"), souligne dans "le financier" le stéréotype qui associe au monde de la finance une autre obédience religieuse: le judaïsme. Selon le groupe auquel appartient le récepteur du message, ses croyances, ses traditions, un même message sera ainsi compris de diverses façons. Ces phénomènes où un unique signal linguistique peut connaître une multiplicité d'interprétations, dont certaines – si elles étaient revendiquées – seraient désapprouvées par le public, constituent ce que nous appelons, dans le contexte de cette thèse, des *dogwhistles*.

Les *dogwhistles* ne sont cependant ici qu'un prétexte. Si le phénomène est intéressant en soi, il est surtout extrêmement atypique, a fortiori si l'on se place dans une tradition de linguistique formelle (et plus encore de linguistique computationnelle). La linguistique formelle est une approche de l'étude des langues naturelles qui tente de les décrire via des représentations

1 Prononcée par Jean-Luc Mélenchon le 29 août 2021, https://www.youtube.com/watch?v=ctoILiwYF3U

supposées suivre les principes du formalisme logique et/ou mathématique. À l'instar de ce que l'on peut trouver dans tout travail de modélisation formelle, comme en physique ou en économie, le but est ici simple : décrire à l'aide d'objets dont on peut parfaitement prédire le comportement celui d'objets pour lesquels on ne le peut pas; décrire en termes "simples" des choses "compliquées". Du fait de la complexité des représentations qu'elle propose, la linguistique formelle – et notamment la sémantique et la pragmatique formelles, dont il est question ici – se concentre généralement sur des phénomènes linguistiques à la fois communs et relativement limités, correspondant à un sous-ensemble bien spécifique des énoncés possibles dans une langue.

L'ambition derrière ce projet tient dans la phrase suivante : tenter d'user des outils de la linguistique formelle – avec lesquels l'auteur a une relative familiarité – pour décrire des phénomènes sur lesquels elle ne se concentre généralement pas.

Dans le cas précis des *dogwhistles*, il semble que l'interprétation qui en est faite par les différentes personnes du public repose non seulement sur les croyances a priori de chacun·e, mais également sur l'identité supposée de l'émetteur du message. Dans la bouche d'un locuteur marqué à l'extrême droite, défenseur du système libéral, l'une des interprétations possibles de (1) devient soudainement plus plausible. D'ailleurs, le terme de *dogwhistle* est parfois plutôt employé pour faire référence à des attitudes ou faits de langue qui, plutôt que de directement transmettre un *message codé* transmettent une *identité* particulière, mais ne le font encore une fois qu'à un sous-ensemble défini de la population, comme une poignée de main secrète.

Le domaine de la linguistique qui se concentre sur le lien entre expression linguistique et appartenance à un groupe est la sociolinguistique. Plus précisément, la propriété d'un signal linguistique qui permet aux auditeurs d'identifier le groupe de l'émetteur de ce signal est généralement appelée *sens social* d'un message. Ce sens social a été largement étudié en sociolinguistique et constitue l'un de ses fondements. Dans son acception la plus contemporaine, on considère généralement que le sens social sert comme vecteur d'identité, ou plus spécifiquement de *personae*, choisies par les locuteurs selon les circonstances. Ainsi, un contexte comme l'entretien d'embauche donnera peut-être lieu à une modulation du langage du locuteur, qui favorisera un certain vocabulaire et une articulation particulière pour renvoyer l'image d'une personne "sérieuse", ou pour cacher une origine sociale qui pourrait desservir.

La théorie avancée dans ce travail est la suivante : l'utilisation de dogwhistles repose sur l'existence d'un système d'interprétation du sens social défini à l'échelle d'un groupe précis ainsi que sur l'existence, au sein de ce groupe, d'une sémantique propre pour l'interprétation de certains items de langage; l'interprétation d'un *dogwhistle* comme marqueur d'appartenance à un groupe par les membres de ce groupe entraîne par la suite son interprétation dans la sémantique du groupe. Les personnes n'appartenant pas au groupe n'ayant pas accès aux interprétations propres au groupe (tant en ce qui concerne le sens social que la sémantique) ignorent l'ambiguïté du signal.

Cette thèse est divisée en 5 parties. La première constitue une introduction naïve au sujet ainsi qu'aux méthodes employées dans ce travail, tandis que la deuxième introduit avec bien plus de détail ces méthodes ainsi que les domaines plus généraux auxquels elles sont rattachées. La troisième partie porte sur une réanalyse plus précise de ce que sont les *dogwhistles* et présente une tentative originale de modélisation formelle du phénomène, qui sert de point de départ à l'élargissement du concept à des cas différents, potentiellement éloignés du domaine de la communication politique. La quatrième partie se concentre sur une tentative de description *algorithmique* du concept de *dogwhistles*, avec en tête la difficile tâche de leur identification dans un corpus de textes. La cinquième et dernière partie, contient, en plus de la conclusion générale du travail, une longue discussion sur les divers raisonnements qui peuvent amener à l'émergence de *dogwhistles*, à leurs réels effets sur le discours, et à leur évolution dans le temps.

PRAGMATIQUE FORMELLE ET SOCIOLINGUISTIQUE

Ce travail, par son objet et ses méthodes, s'inscrit avant tout dans la tradition de la pragmatique formelle. La pragmatique est généralement définie comme étant la partie de l'analyse linguistique qui se concentre sur l'étude du sens des *énoncés*, c'est-à-dire des messages linguistiques dans le contexte dans lequel on les produit. Du fait du phénomène précis autour duquel tourne ce travail, des considérations issues d'autres domaines de l'étude linguistique ont été mobilisées, en particulier provenant de la sociolinguistique.

Si les traditions logicienne et formaliste sont depuis longtemps présentes et reconnues en sémantique, les questions considérées comme relevant de la pragmatique n'ont que récemment fait l'objet de formalisations similaires. Depuis au moins Wittgenstein (1953), la notion de *jeu de langgage* propose de voir le sens des mots d'une langue non comme étant leur dénotation (c'est-à-dire les objets auxquels ces mots font référence dans le monde), mais comme étant leur usage en contexte, dans le cadre d'interactions ayant d'autres objectifs que la simple description de ce qui est. Avec Lewis (1969), la notion de *jeu de langage* est comprise dans une dimension mathématique, et c'est ce travail qu'on place généralement à l'origine de la représentation des processus de pragmatique faisant usage des concepts issus de la théorie des jeux mathématiques.

La théorie des jeux est une branche des mathématiques utilisée pour décrire les situations d'interactions entre agents (ou "joueurs") telles que l'issue de l'interaction pour chacun de ces agents dépendra de ses propres choix ainsi que de ceux des autres agents. Dans la tradition en pragmatique formelle, les dialogues et l'interprétation des messages est parfois représentée par le concept de jeux dits *de signaux* ('signaling games'), une formalisation d'une situation de communication. Dans un jeu de signaux, un agent dispose d'une information et souhaite la communiquer à un autre agent. Pour ce faire, l'agent 1 dispose d'un certain nombre de

signaux. Le but du jeu est, pour l'agent 1, de choisir le signal qui permettra le plus efficacement de communiquer à l'agent 2 l'information qui lui est inconnue.

Cette situation abstraite ainsi que la stratégie dite de *meilleure réponse itérée* ('iterated best response') forment la base de la famille de modèles appelés Rational Speech Act (RSA). Ces modèles ont notamment servi à la description de plusieurs phénomènes observés en pragmatique.

Les *implicatures scalaires* sont un tel phénomène. En quelques mots : on parle d'implicature scalaire lorsque l'existence d'une alternative plus précise à un message implique que cette alternative soit fausse (ou du moins que le locuteur la pense fausse) si elle n'est pas employée. Par exemple, la phrase en (2a) implique généralement que celle en (2b) est fausse, alors même que les deux situations sont en fait compatibles (et que, d'ailleurs, si (2b) est vraie, alors (2a) l'est nécessairement).

(2) a. L'accident a fait trois blessés.

b. L'accident a fait quatre blessés.

Une formalisation en utilisant le cadre RSA permet de montrer comment cette assymétrie sémantique permet l'émergence de l'implicature. Toute situation opposant deux alternatives dont l'une a une référence qui est un sous-ensemble strict de la référence de l'autre entraînera en RSA l'effet d'implicature scalaire, et cette alternative plus précise sera, après plusieurs cycles du jeu, considérée comme fausse (ou du moins très peu probable) si le signal qui ne lui est pas directement associé est employé.

Dans Burnett (2017, 2019), le cadre RSA est adapté à la descrption des processus d'interprétation du sens social. Plutôt que de rattacher à chaque message une interprétation sémantique, on y attache cette fois un sens social, sous la forme d'un ensemble de propriétés définitoires de personae qui seront rattachées à une variante sociolinguistique plutôt qu'à une autre. En utilisant les mêmes mécanismes que le RSA standard, on peut ainsi construire les Social Meaning Games (SMG), qui traitent le sens social de la même façon que l'on pourrait traiter du sens sémantique standard et parviennent à généraliser le processus de précision du sens qu'on observe par exemple dans les implicatures scalaires pour décrire une classe de phénomènes a priori indépendante.

Les modèles RSA et SMG sont les briques fondamentales du modèle original qui sera proposé dans la thèse, les Dog Whistle Games (DWG).

À bien des égards, on peut voir les DWG comme la fusion du cadre RSA standard et des SMG, mais cette fusion ne se fait pas sans heurt. L'existence d'un sens social et d'un sens sémantique dans chaque signal et l'utilisation de l'un pour le calcul de l'autre mènent à des situations irrégulières lors de l'interprétation des agents de notre modèle. Le modèle permet néanmoins de capturer de nombreuses caractéristiques des dogwhistles, et notamment la possibilité d'être plausiblement niés par les personnes les employant.

Le modèle DWG est par ailleurs adaptable à la description de phénomènes connexes mais néanmoins distincts, tels que l'utilisation de messages à double sens dans la production artistique (nous nous concentrons sur deux études de cas : la poésie de Walt Whitman et la chanson de Sufjan Stevens), ou l'utilisation d'expressions dont la référence, peu définie, donne lieu à des myriades d'interprétations, parfois à l'échelle des individus eux-mêmes.

APPROCHE COMPUTATIONNELLE

Une caractéristique importante des dogwhistles est leur aspect caché. Conçus pour passer inaperçus dans un discours, ils sont de ce fait difficiles à débusquer et leur identification repose ultimement sur l'avis personnel de l'investigateur. Dans l'optique de donner une base moins individuelle à ce processus d'identification, ce travail a tenté d'employer des concepts issus de la sémantique distributionnelle et de la linguistique computationnelle pour tenter de donner une définition des dogwhistles qui soit opérationnelle pour leur identification automatique.

Plutôt que de voir la tâche d'identification comme une tâche classique d'apprentissage automatique cependant, ce travail propose de se servir de représentations communes en sémantique computationnelle, et notamment des *plongements de mots* ('word embeddings'), comme d'un outil d'exploration des données. Là où la notion de biais dans les systèmes d'intelligence artificielle s'avère catastrophique pour un nombre important d'applications de ces systèmes, elle devient dans le cadre de ce travail une propriété des modèles permettant d'inférer certaines des propriétés des corpus de texte employés qui seraient autrement plus difficilement accessibles.

Dans le cas de l'étude des dogwhistles, nous avons ici tenté de comparer des espaces vectoriels de mots associés à des communautés différentes pour tenter d'observer si des mots soupçonnés de porter des significations différentes en fonction de ces communautés reflétaient effectivement ces différences. Nos cas d'étude sont les corpus de débats de l'Assemblée Nationale portant sur la légalisation du mariage homosexuel ainsi que sur le PAct Civil de Solidarité (PACS).

LES DOGWHISTLES : POURQUOI ?

L'ultime chapitre de la thèse revient sur les apports et accomplissements du travail ainsi que sur ses limites, et propose une réflexion supplémentaire sur l'intérêt de l'utilisation des dogwhistles dans le débat public, qui va potentiellement plus loin que l'horizon d'une élection prochaine, et soulignant le lien profond qui existe entre dogwhistles et propagande.

Il semble évident que la linguistique seule ne suffit pas à expliquer les phénomènes auxquels nous nous sommes intéressés ici, et que les approches formelles et computationnelles en particulier, si leur apport est certain, trouvent vite leurs limites dès lors que les phénomènes que

l'on souhaite étudier doivent être compris à la lumière non d'une vision du langage comme capacité cognitive, mais comme objet culturel et social.

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

Dénigot, Quentin and Heather Burnett (June 2020). "Dogwhistles as Identity-based interpretative variation." In: *Proceedings of the Probability and Meaning Conference (PaM 2020)*. Gothenburg: Association for Computational Linguistics, pp. 17–25. URL: https://aclanthology.org/2020.pam-1.3.

— (2021). "Using Word Embeddings to Uncover Discourses." In: *Proceedings of the Society for Computation in Linguistics*. Vol. 4. DOI: https://doi.org/10.7275/t4y8-z343. URL: https://scholarworks.umass.edu/scil/vol4/iss1/28.

ACKNOWLEDGEMENTS

Several people close to my heart study some flavor of history and/or philosophy of science. These people tend to have strange obsessions, such as looking at dissertation acknowledgements to retrace the academic genealogy of persons of interest. In the very unlikely event that a student of history and philosophy of science from the future reads this, I have tried to make the information useful to them clear enough, and have written it in English.

Custom dictates that the first person to thank be the advisor, custom sometimes dictates right. There is no overstating the importance of the part Heather Burnett played in this work. She has been an inspiration from start to finish, and without her input, this dissertation would be very different. In fact, without her steadfast support, it probably wouldn't exist. There is no thanking her enough, as a teacher, as an advisor, as a mentor, as a friend.

This work would not have seen the light of day either if it weren't for the involvement of Barbara Hemforth, who, undeterred by the fact of not knowing anything about me, accepted to associate her name to this work and believed in my ability to not mess up too much. Although she did not remain an advisor of this thesis, her initial input and involvement despite the subject being relatively far from her usual interests were both very useful and greatly appreciated.

Olivier Bonami and Denis Paperno both showed great benevolence and kindness with regard to my work and person in filling their role as my *comité de suivi*. Their advice was invaluable and my thanks are extended to them as well.

Many thanks to Nicholas Asher, Jonathan Ginzburg and Denis Paperno (again) for accepting to be part of my thesis jury; many thanks in particular to Elin McCready and Jennifer Saul for accepting to be *rapporteures* for this dissertation. Your works have both been hugely inspiring and enlightening in the entire process of researching and writing the thesis. I hope I have done it justice and I hope maybe someday to be as inspiring to anyone's work as you both were to mine.

Acknowledgements also go to all the people that have in some way or other participated in this work in the form of general discussion and/or proofreading. Many thanks in particular to Jamie Findlay, Marie Candito and Asad Sayeed for their much appreciated inputs and remarks, and to Valentin De Bortoli and Théis Bazin for their patience and assistance in all things math.

Merci également au Laboratoire de Linguistique Formelle tout entier, qu'il s'agisse de la présidence, des chercheureuses et maître·sse·s de conférence, des personnels administratifs, des ingénieurs de recherche... C'est cette structure toute entière qui fait que je garderai sans nul doute de ma vie de thésard le souvenir d'un moment de découvertes et d'amitiés, et

pas seulement de travail et d'isolement. Merci également à l'ED 622, sa présidence et ses personnels, qui ont en toute circonstance su répondre à mes questions et incertitudes.

Merci aux collègues doctorant·e·s et post-docs, dont la présence a su illuminer mes journées. Merci notamment à Aixiu, Yiming, Pengye, Saida, Maria, Bingzhi, Peijia, Justine, Rahma, Suzanne, Angélique, Gabrielle, Gabriel, Céline, Juliette, Zulipiye, Beatrice, Zhanglin, Marie, Cécile, Yanis, Julie et Alex. Vous rencontrer a été une chance, vous fréquenter un plaisir, nul doute que nos chemins se recroiseront un jour, sur les bancs d'une université ou ailleurs.

Merci bien sûr aux ami·e·s-qui-ne-sont-pas-par-ailleurs-des-collègues, sans qui tout est un peu plus nul. Merci à Margaux, Laura-May, Alice, Manon B, Manon V, Clément, Valentin, Théis, Jeanne, Corentin, Antoine, Sarah, Audrey, Cécile, Başak, Léonard, Emma, Lorenzo et Pierre-Léo.

Merci également (et mes excuses) à toustes celleux que j'oublie de nommer alors que j'écris ces lignes.

Merci bien sûr à ma famille : à mon père et ma soeur pour me soutenir, toujours ; à ma mère pour m'avoir soutenu, toujours.

Merci enfin, évidemment, à Margo, pour tous les jours passés et tous ceux à venir.

Pas merci à SARS-CoV-2.

CONTENTS

LIST OF FIGURES

ACRONYMS

CDA Critical Discourse Analysis

CP Cooperative Principle

DS Distributional Semantics

DWG Dogwhistle Game

GCI Generalized Conversational Implicature

GT Game Theory

IBR Iterated Best Response

MEG Message Exchange Game

NLP Natural Language Processing

RSA Rational Speech Act

SMG Social Meaning Game

VSM Vector Space Model

WE Word Embeddings

Part I

AN INTRODUCTION

In this introductory part, we will discuss many of the fundamental concepts and ideas behind this work and its link to various branches of linguistics as well as presenting the key questions that we will try to answer in the rest of the work.

1

INTRODUCTION

"Style is the man himself"[1]

— Georges-Louis Leclerc, comte de Buffon (Buffon, 1753)

"I change my style maybe every month. I'm, like, punk one month, ghetto fab the next, classy the next. I'm just young and finding out who I am."

— Kylie Kristen Jenner[2]

1.1 THE PROBLEM

Variation and style

In his *Discours sur le Style*, Buffon, then recently made a member of the very exclusive *Académie Française*, underlines how it is not facts, knowledge or novelty in a text that achieve posterity, but *style*, for style is the true mark of the author. Buffon's thoughts do not necessarily limit themselves to literary writing; in fact it is when discussing a very different subject that Victor Klemperer invokes that same quote in *LTI*, his review of language as it was used in Nazi Germany:

"*Le style, c'est l'homme*; no matter how deceitful a man's claims might be, the style of their language exposes their true self"[3]

— Victor Klemperer (Klemperer, 1947)

In both cases, *style*, left undefined but understood as a manner in which to use language to articulate thought, is tell-tale of *identity* and similarly *identity* dictates *style*; there is an imprint of each on the other.

But what about that Kylie Jenner quote? It seems to ascribe a very different meaning to *style*, likely in the sense of *clothing style*. Not only that, Jenner seems to imply that style is not constant. She even claims to have herself changed styles many times. Surely then, *style* can't be "the man himself", either that or there is no such thing as "the man himself", and there is just a succession of styles and identities.

1 "Le style est l'homme même."
2 I am assuming this appeared on her instagram page, though I could not trace the source of the quote.
3 "*Le style, c'est l'homme*; die Aussagen eines Menschen mögen verlogen sein - im Stil seiner Sprache liegt sein Wesen hülenlos offen."

So what is it? Is style the great revealer of true identity, or are people adorning themselves in one way or other according to personal preferences? Oddly enough, both positions can and will be defended here.

See, conveniently, this dissertation will focus on uses of language for the construction of identity. But there's a catch. The title of the dissertation mentions "*conflicting* identities", implying there can be more than one. Let's look at the linguistic display of identity for a moment.

From a linguistics perspective, the mapping between identity and manner of speech is both trivial and oddly evasive. For example, it goes without saying that phonological variation can be a reasonably reliable indicator of geographical origin through *accents*. But *accents* are not *style*, they do not emanate from a conscious choice[4]. Regional variation is however favored by a number of parameters, and while *accents* seem most of the time not to result from conscious decision, there are pressures (both internal and external to the speakers) that will lead to the preferred use of one variant over another and its possible subsequent establishment as a community's favorite pronunciation. The now classic Labov (1963) shows how the pronunciation of diphtongs /ai/ and /aʊ/ by locals on Martha's Vineyard has seen a rise in the island-specific centralization pattern in the decades preceding the study in spite of a growing number (and one could assume a growing influence) of outsiders on the island. One reason for the apparent phonological conservatism of Vineyarders presented in Labov (1963) lies in resentment towards outsiders buying land on the island, which translates into the pronunciation of /ai/ and /aʊ/ as [əi] and [əʊ] being markers of "vineyardness". Maintaining the identity of the group seems to have had a key influence in the presence of these phonological features.

Although the centralization of diphtongs on Martha's Vineyard seems to be largely unconscious to users of that variety of English, it seems its value as an identity marker is a good explanation for its presence; centralizing diphtongs makes you *sound like* a Vineyarder. While this in itself would not come under a standard definition of *style*, we're not too far from it, because once such variables have been uncovered by speakers, what's stopping them to use those to *sound like* someone from a certain group?

Work on variation has progressed a lot since Labov (1963), and it is now generally accepted that variation is not merely a marker of intrinsic social properties of a speaker (such as their belonging to a given social group), but can also vary *at the scale of the idiolect of an individual* to temporarily mark their belonging to a group, a role, a social class, or in the hopes of displaying specific properties to their listeners[5].

A key insight of variationist approaches to sociolinguistics can be summed up as follows: each linguistic signal contains information about the identity of this signal's source. That information that's added to the signal and is dependent upon the source, would traditionally be considered to be "noise". But it has regularities, it does contain *information*. The information about the source of a signal carried by a linguistic signal is sometimes called *social meaning*,

4 Or do they? See Trudgill (1986).

5 This will be taken up again in chapter Chapter 3, section 3.1.1; for a concise and complete review of the study of variation in sociolinguistics, see Eckert (2012).

to be differentiated from what could be called *semantic meaning*, which would refer to the content of the message itself (traditionnally expressed in terms of truth-conditions following Tarski's semantic theory of truth). We will come back to the notion of social meaning and how it can relate to more traditional topics in pragmatics in part ii. How social meaning actually emerges and how its structure can change over time to trigger new interpretations is an open question, but we will see that it can be useful to think of it as a supplementary layer of meaning that does not necessarily require a very different treatment from semantic meaning and that existing frameworks for the derivation of implicatures can be useful in that endeavor. This question and more will be explored in chapter Chapter 3, in particular using the work presented in Burnett (2017, 2019).

Of course, variation is not limited to *phonological* variation, we can also mention cases of *syntactic* variation, variations in *register*, *lexical* variation... Similarly, *style* refers to a set of properties displayed by individuals, among which are linguistic variants (Eckert, 1988), but also, e. g., clothing. We will use that term, *style*, specifically when the variation is observed at the scale of an individual speaker. The word *persona* is sometimes used to refer to styles displayed by an individual, or rather to sets of *properties* that an individual wishes to display through the use of a specific style[6] , and it is the way we are going to use it in the context of this work. To sum it up, *style* is the use of particular *variants* indexing specific sets of *properties* through *social meaning*. These sets of properties are called *personae*.

Finally, the kind of variation that we will be focusing on here is mostly *lexical*, although we hope this work can provide insights on the study of other forms of variation, just as much as it has found insights in their study.

Variation and discourse

Why lexical variation? Surely there are many parameters other than style that come into account when choosing which words to use.

This is true, but the reason we will be focusing on lexical variation is that we will in particular be interested in a phenomenon often called *dogwhistle politics*, which can be described as a "way of sending a message to certain potential supporters in such a way as to make it inaudible to others whom it might alienate or deniable for still others who would find any explicit appeal along those lines offensive" (Goodin and Saward, 2005). Although this cursory definition does not exclude any form of message-sending, the term is most often used to either refer to lexical units or broader discourse topics. We will argue throughout this thesis that the phenomenon behind the interpretative variability at the heart of *dogwhistle* phenomena is linked to a display of group membership through the use of certain lexical items. More specif-

6 One might also have heard the term *code-switching* or *code-mixing*, which generally refers to stylistic and linguistic variation (notably through the use of a variety of dialects or registers) at the scale of a single discourse or interaction. That intra-discursive/intra-sentential variation will not be discussed here.

ically, we will argue that those terms used as *dogwhistles* can be seen as playing on a *semantic* variability that is intrinsically linked to identity display in the language.

To illustrate this notion, let's look at one of the most cited examples in the literature:

(3) a. Yet there's power, *wonder-working power*, in the goodness and idealism and faith of the American people.

b. Yet there's power, *Christian power*, in the goodness and idealism and faith of the American people.

Emphases mine

(3a) is a sentence pronounced by then POTUS George W. Bush in Bush (2003). The phrase "*wonder-working power*" is an excerpt from a famous hymn, *There is Power in the Blood*. The usual analysis of this example is that this specific phrase is used as a signal to a subset of Bush's supporters – evangelical voters – whose interpretation of the phrase is analyzed as (3b). It works like a verbal secret handshake, the group of evangelicals recognize Bush as one of their own, the others ignore the signal.

Using this seemingly neutral (albeit strange) phrasing allows George W. Bush to communicate his belonging to a specific community (evangelicals) and at the same time use their group-specific interpretations to communicate a supplementary layer of meaning, while avoiding more openly Christian phrasings such as the one in (3b), which would alienate some potential voters, for whom politics and religion should not mingle.

We will be coming back to that specific example and more in the dissertation, but it underlines our analysis of dogwhistles fairly well: some lexical items and turns of phrases are specific to a group but understandable in some way by users of the language outside that group; these items act as a social signal to members of that group, whose interpretation of the items thus follows a group-specific pattern; listeners outside of the group do not follow that pattern of interpretation. Because the speaker has signalled their belonging to a specific group, members of that group will interpret their messages following the norms of their sociolect, and members outside of that group will ignore the signal and interpret the message in their own sociolect, making different inferences.

At least, this is one theory about dogwhistles, but we will see that even though the literature more or less agrees on the informal definition of dogwhistles we've given and on analyses similar to this one, observed cases are hardly as straightforward as this. Reevaluations of the effects of dogwhistles will notably be tackled in Chapter 4 and Chapter 9.

Being largely associated with political discourse, *dogwhistles* are deeply linked to the notion of *discourse* itself, which is understood here as socially-situated language use. Terms that can be labeled *dogwhistles* thus acquire their dogwhistley traits in specific social interactions, and when discussing specific subjects, at specific points in time, with a specific audience.

A corollary to the very basic definition we have given of dogwhistles above is that they have at least two interpretations, including one that is not "alienating". We will argue that at the scale of a given discourse, it is possible that some terms display semantic variability that correlates with the identity of the person using them, being thus imbued with some key properties of dogwhistles and leading us to more functional definitions.

That *non-alienating* part of dogwhistles is key to how and why they are used. The existence of dogwhistles as they are usually defined implies the existence of communities with conflicting interests on the one hand, and the necessity for a given agent to gather the support of several of those communities on another. One hypothesis that will be put forward in this work is that dogwhistles allow such an agent to gather the support of both communities by appearing to be in keeping with both communities' norms and preferences. The underlying idea is that the agent in question uses social meaning cues to appear to be part of both communities at once; that the agent communicates simultaneously different identities – seemingly *conflicting* identities.

1.2 METHODOLOGIES

Formal and mathematical approaches represent a significant part of the work presented here, but all formal descriptions rely on the contribution of non-formal approaches in the preliminary analysis and construction of the concepts; no study of a linguistic phenomenon would be complete that relies exclusively on formal definitions. And in any case:

> *"If we want to understand nature [. . .] then we must use all ideas, all methods, and not just a small selection of them."*
>
> — Paul Feyerabend (Feyerabend, 1975)

This is one way in which this work could be qualified as "interdisciplinary": it will rely on literature in many fields, and has stemmed from conversations with people from many horizons.

That being said, we will rely on two key approaches to tackle the issue.

Formal approaches

Matters of variation, identity construction and social meaning in general are, we argue, central to the human linguistic and social experience. While these phenomena have been extensively studied in sociolinguistics, they have been left aside by formal approaches to linguistic description. A first interrogation as to the reasons behind this lacking would be to simply say that formal semantics is mostly referential and interested in literal meanings, with little regard for context of utterance, but that criticism of the field would be a bit unfair.

While it is true that formal approaches to the study of meaning have very largely focused on the *content* of messages – what we incidentally called *semantic meaning* above – and have tried to define it first in terms of reference to objects in the world, it has been very clear from the start that a purely referential approach would be insufficient for many (maybe most) sentences (Frege, 1892). For example, what do we do of expressions that share a reference only for a given amount of time, like "Biden" and "POTUS"? What do we do of the reference of deictics such as "I" or "here"? Authors like Russell and Frege already had those questions in mind, and there have been several attempts at implementing notions of context into natural language semantics.

That being said, the notion of meaning in context is usually more readily associated with *pragmatics*. The actual difference between *semantics* and *pragmatics* is far from being clear (Korta and Perry, 2020). One way of describing it would be to say that the objects of study of semantics are *sentences* while the objects of study of pragmatics are *utterances*, meaning that even when semantics takes an interest in context (e. g.possible world semantics, dynamic semantics), it does not necessarily focus on the meaning of any particular utterance. The semantics of a sentence in that way can be understood as the set of all possible meanings of that sentence once it is uttered. Pragmatics, on the other hand, focuses on the meaning of the actual utterance.

A naïve reading of this would subsume pragmatics to applied semantics, but that would again be unfair. Indeed, there is at least one aspect of meaning in context that is generally thought to be outside the realm of semantics, the very multi-faceted case of *implicatures* (systematically studied for the first time in Grice, 1967). Implicatures arise from the fact that what the sentence *says* is not necessarily what the speaker *means*, and in fact that the correct interpretation of an utterance can have little to do with its actual content. As a concept, implicatures allow us to redefine pragmatics as also being the study of speaker intention, or more precisely the inference of speaker intention from a given message.

A lot of *meaning* as it is understood by users of language has in fact more to do with deciding what *is meant by the speaker* than with *what the sentence means*.

As an informal approach to language meaning *in use* and general meaning construction, pragmatics has been very insightful and its main authors now constitute an important part of a linguistics student's syllabus (Grice, Austin, Searle, Lewis, etc.), but actual attempts at formalizing these insights are actually fairly recent, and some time was needed before mathematical tools appropriate for the description of interactions and message exchanges were actually applied to the study of language use between agents[7]. While the late Wittgenstein is probably to be credited for putting at the forefront the term *language game*, it is more likely David Lewis that we have to thank for the more precise formulations of what these games might be, formulations that in turn led to formal frameworks for the study of pragmatics

7 "The problems of pragmatics have been treated informally by philosophers in the ordinary language tradition, and by some linguists, but logicians and philosophers of a formalistic frame of mind have generally ignored pragmatic problems." Stalnaker (1970)

characterized in game-theoretic terms, a nowadays fairly popular way to describe part of the meaning that remained elusive to more traditional semantics.

What we argue here is that the methods and tools that have been used to handle, e. g., implicatures and in general *meaning beyond the sentence*, can also be taken as a basis to describe the phenomena that we have been discussing earlier. In the line of existing work on similar subjects (Burnett, 2017, 2019), we will try to extend the horizons of the game-theoretic pragmatics programme and attempt to use this framework to elucidate a greater array of phenomena related to language interpretation.

But let's take a step back: why would we want to give a formal account of these phenomena? Aren't the informal accounts sufficient already?

Formal approaches provide a way of representing phenomena using language that purports to be perfectly defined and provides objects that are predictable to the point that highly precise hypotheses can be formulated about them. Using formal languages like mathematics to describe linguistic and other natural phenomena can be thought of as using objects that we understand to describe objets that we don't, with the hope of better understanding how the latter work thanks to knowing how the former work[8]. In that respect, they have a lot to offer both from theoretical and empirical standpoints, but the mistake should not be made of conflating the formal objects that we use for our representations with the natural objects we wish to describe.

The key word here is *representation*. Formal models of phenomena give us a possible representation of those phenomena, and the subsequent formal proofs about the behavior of those formal models are an artifact of the formal language in which they are represented. Only if the deterministic behavior of our formal models is congruent with the observed behavior of their real world counterparts are the models themselves relevant and useful for real-world use. But whether or not they happen to be, they can give us insights on the real world objects that might not have been reached otherwise, forcing us to be more specific in our definitions.

Attempting to build a formal model of a phenomenon forces one to consider the many interpretations allowed by the informal description of that phenomenon. In other words, attempting to formally define something like dogwhistles should lead us towards a more precise, less equivocal definition of what they are, and should underline the need, if it exists, to divide the concept into several categories.

Computational approaches

Beyond definitions, the issue of data also arises. Over the last decade, with the advent of social media platforms, people have created incredible amounts of content, and a wide array of phenomena in social sciences and the humanities are increasingly being studied through

8 That idea and its formulation are not mine, I believe I read them first in the introduction of a book (I believe it was a book by András Kornai). No matter how long I have been looking for the original quote, I could not find it.

the lens of the analysis of large datasets. As the emergence and establishment of fields like digital humanities or computational social sciences suggest, *text* is increasingly being seen as *data*. This shift in practices is well described in Nguyen (2017).

Not only that, but the much more systematic digitization of public archives, many of which are publicly available, also appears to be a treasure trove of text data waiting to be analyzed.

In this work, we will attempt to address some of the shortcomings that appear in traditional analyses of dogwhistles using the insights of computational social sciences. The issue that was a starting point in this endeavor is the fact that the same example dogwhistles are re-used many times across works. (3) has for example been used as a textbook example several times (Albertson, 2015; Henderson and McCready, 2018, 2019b; Saul, 2018a), and a few others (e. g.*inner cities*) often come up. While having these stable interpretations of clearly identified occurrences is good for exposition purposes, it also feels like these examples are nowadays a bit old, and might not be very representative of dogwhistles "in the wild". They also happen to mostly apply to US contexts.

But with little information about how frequent this phenomenon might be and each reported occurrence relying heavily on the personal impressions of the commentator, we are at risk of both using not-so-representative examples for the basis of our theories and being accused of cherry-picking.

Using the insights from our formal analysis and from non-formal accounts of the phenomenon, we have tried to find ways to describe dogwhistles in *computational terms*, i. e.using what we have found as the basis for an algorithmic approach that would theoretically allow us to find dogwhistles in text automatically. Although this endeavor has had mitigated success, it has led us to use tools taken from distributional semantics and machine learning and apply them to topics which, to our knowledge, they had never been applied to before, giving us insights both on the concept of dogwhistle detection itself and on the properties and limitations of those methods, notably when applied to corpora that would typically be used to try and answer the questions that interest us.

It is important to note that we are adopting a posture that is very different from the typical posture adopted when using these methods. Computational methods, and especially those found in fields like machine learning and deep learning, are often used with a specific task in mind. These methods are subsequently validated on success in accomplishing that task on some data, the assumption being that the performance in training will generalize (at least to a point) when deployed in the real world. Here, we are not task-focused, the methods that we will use will help us not in a specific task, but as a way to bring new insights on the data directly, as a way to represent the data that we're interested in. We will try to interpret the results that we get in this light, and not according to the usual metrics for success and accuracy[9].

9 See Nguyen (2017, pp. 198-9) for interesting insights regarding the use of computational methods in computer science, focused on accuracy of prediction and success in application tasks, and their use in social science and humanities, focused on interpretability of results and theory-building.

1.3 RESEARCH QUESTIONS AND OUTLINE

Dogwhistles are a fairly strange object of study from a linguistics point of view because they go way beyond the mere linguistic experience. Thinking about them in terms of undercover diplays of identity leads to studying the works on identity and social meaning in sociolinguistics, but their political nature questions some of the more familiar linguistics approaches. Specifically, as a topic, they force us to study interpretation and communication processes under *non-cooperative* conditions, counter to what many works focused on linguistic meaning have done.

This brief introduction to the topics of this thesis underlines the main thing that I have attempted to do during these three years: use the tools of formal linguistics – with which I was somewhat familiar – to describe phenomena that were largely ignored by formal linguistics – which I found rather interesting. Computational considerations emerged along the way.

QUESTION: WHAT ARE DOGWHISTLES? The first research question that we focused on was the need and possibility of giving a formal definition to dogwhistles. Part iii will provide both a collection of definitions found in the literature and our attempt at formalizing the concept. In this part we will present the game-theoretic model on which we have worked to give dogwhistles a new representation that actually relies on existing works in pragmatics and cognitive science.

Indeed, the work presented here is largely based on extensions of the RSA family of models, first presented in Frank and Goodman (2012) and now applied with success to several areas of pragmatics, notably scalar implicatures.

We will also see that there are situations analogous to dogwhistles that can arise in non-political contexts, specifically in art, and where the covertness of dogwhistles is not necessarily as relevant as their apparent vagueness.

QUESTION: CAN MY COMPUTER FIND DOGWHISTLES? As said earlier, we now have access to immense amounts of data on social media. Due to the high frequency of exchanges and the sheer volume of content created by users, it is possible that dogwhistles appear and disappear faster than ever, and it is possible that we miss many examples.

In part iv, we will focus on attempts at text data analysis that could lead to automatically (or at least reliably) uncover dogwhistles and more general group-based semantic variation. Using publicly available corpora of political debates, we try to see how some measures of semantic similarity correlate with the identity of speakers.

QUESTION: WHY AND HOW ARE DOGWHISTLES USED? Dogwhistles are highly contextual in nature and should in theory be rather short-lived (once everyone is in on the secondary

meaning, the dogwhistle ceases to be used as such). In Chapter 4 and Chapter 9, we have tried to leave the fixed synchronic representations of the formal model to tackle the question of the evolution of dogwhistles, which is not a clear matter. In the process, we will also discuss the reasons that can push one to use such seemingly inefficient communication strategies, and what their effect on discourse can be, notably in the long term.

Before anything, however, part ii will focus on a more in-depth introduction to the concepts of *game-theoretic pragmatics* and *social meaning*, which were the two starting points behind the project. These concepts stem from very different traditions, and that part will therefore also focus on existing examples of attempts to join the two. In many ways this part acts as a more specific introduction than the one presented thus far, in that it will notably present the main formal tools that we will be using in the rest of the work.

Part II

BACKGROUND

This part focuses on an in-depth explanation of why this work is first and foremost a work in *pragmatics*, before presenting the concept of social meaning and how it relates to the study of pragmatics.

2

PRAGMATICS AND INTERPRETATION

"What do you mean?"

— Justin Bieber, singer-songwriter, probably thinking about pragmatics (Bieber, Boyd, and Levy, 2015)

"Language serves many important purposes besides those of scientific inquiry; we can know perfectly well what an expression means [...] without knowing its analysis..."

— Paul Grice reminding everyone that formal analyses are not, in fact, a pre-requisite to comprehension. (Grice, 1967)

A friend of mine[1] once told me something along the lines of "I read the Wikipedia entry for linguistics, they mention all the subfields, and it was all pretty clear to me what they all were. Except for pragmatics. I understood nothing. What on Earth is it about?". I tried to answer and realized that it was not clear at all what it is that pragmatics is. Because this work is a work in pragmatics (or at least related to pragmatics), it might be useful/desirable to have a reminder of what we mean when we say "pragmatics" in linguistics.

More specifically, all along the chapter, I will try to progress towards the idea of *language* being represented as a *game*. When pragmatics and the branch of mathematics known as game theory collide, we get *game-theoretic pragmatics*, the tradition in linguistics and cognitive science in which this work (especially part iii) finds itself.

2.1 FROM SEMANTICS TO PRAGMATICS

It is generally accepted among people that a linguistic signal, whether it is spoken/heard or signed/seen, is not *just* a sensory input, it also has *meaning*. The study of linguistic meaning is the focus of $\frac{\text{semantics}}{\text{pragmatics}}$[2].

Why two fields for one object of study? Although we will see that the distinction between the two is not easy, there are good epistemological reasons to make it. And in fact, whenever pragmatics – which is here our main concern – is defined, it is usually in contrast with semantics. For example, the many definitions presented in Korta and Perry (2020) define pragmatics in this way. This "definition by comparison" underlines both the chronological relationship between the two fields (semantics, as an academic object of study, has existed for a longer

1 Clément

2 This clever way of saying two things at once is inspired by the Ariekan language, as presented in Miéville (2011).

time than pragmatics has) and the two different conceptions of "meaning" put forward by the two fields.

In the case of semantics, and at least since Frege (1892), a distinction is made between what constitutes *sense* and what constitutes *reference*, and taking this as a starting point, a lot of semantics has initially consisted in constructing a theory of *reference*. The reference of a lexical unit is the object or set of objects that it *refers to* in the world[3]. A theory based on reference can allow us to account for a number of sentences, and in particular all those sentences that describe the world, often in the form of *propositions*. One question that has been important to semanticists, inspired by logicians and philosophers, is the following: what makes a sentence (like (4)) true or false?

(4) Dogs are beautiful.

Well, in the case of (4), the sentence would be true in case the set of objects that "Dogs" refers to (the set of all dogs) is a subset of the set of objects that have the property referred to by the word "beautiful"[4]. Once we've established the reference of all words in a proposition – whether that reference is a set of objects ("Dogs"), a property ("beautiful", also a set of objects), a relation between sets ("are") or anything else –, then the magic of *compositionality* – which is the principle saying that the meaning of sentences is a composition of the meaning of its parts – gives us the *truth conditions* of that proposition. This certainly is not enough to give a picture of the whole of what is meant by "meaning", but it is a start. A famous limitation to a reference-only theory of meaning is already found in Frege (1892), with the example of Hesperus and Phosphorus, the Evening Star and the Morning Star, which both happen to refer to the same celestial body (the planet Venus). A purely referential theory of meaning could not account for the apparent difference in meaning between the two examples in (5), since they have the same reference:

(5) a. Hesperus is Hesperus.

b. Hesperus is Phosphorus.

Those two sentences are however different: the first one is trivial and uninteresting, while the second one is possibly a huge revelation about the nature of Hesperus/Phosphorus. This difference is part of what Frege calls *sense*. It feels more elusive than reference and has been left aside by many semanticists with formal inclinations.

What are those *truth conditions* that a referential approach allows us to find? We can define them as the conditions under which a given sentence would be considered to be true. Though this seems at odds with our intuition, it has been argued that knowledge of truth conditions is equivalent to knowledge of a language, more specifically, that understanding the

3 Their *extension*.

4 Which is the case, this is a true proposition.

circumstances under which any sentence of a language is true is equivalent to understanding that language (Davidson, 1968). While this position is not shared by many, it remains true that truth conditions and theories of reference have remained important in the study of semantics[5].

Works in the line of Saussure's have put the emphasis on the study of *langue* in a synchronic manner, as a rigorously defined autonomous object, and have generally treated the meaning of any message conveyed using any *langue* in the same way.

> *"The unique and true object of linguistics is langue, envisioned in itself and for itself."*[6]
>
> — de Saussure, 1971

For example, in Hjelmslev (1957), we are presented a programme for the study of natural language semantics that first focuses on the commutation of different units of language with or without change in the content, in a similar vein to what was done in phonology using minimal pairs, and in the hope of looking at semantics as a *structure*, or "autonomous entity of internal dependencies"[7], a necessary step for scientific inquiry[8]. While it is acknowledged that this would not be enough to account for the entirety of what we generally understand by *meaning*, it is seen as a necesary first step and in fact the only way it can be done scientifically. Looking back at the previous Saussure quote, any study of context (or more generally of anything that falls oustide the scope of *langue*) is not the object of linguistics.

In keeping with earlier and contemporary literature on the matter, Saussure and early academic linguists focused on the idea that a *speaker* had in mind a *concept* that they wished to convey to a *listener* through the use of *langue*. This early approach sees meaning as being, as a whole, encoded in the linguistic signal, and retrievable thereof. Truth conditional and reference-oriented semantics has focused on a similar idea of meaning, with each sentence encoding its truth conditions, therefore its meaning. Anything that goes beyond a study of the *reference* of the linguistic objects and includes data outside the language itself goes beyond the study of semantics. This position can be found in e. g.Montague (1968) and Morris (1938):

> *" Syntax is concerned solely with relations between linguistic expressions; semantics with relations between expressions and the objects to which they refer; and pragmatics with relations among expressions, the objects to which they refer, and the users or contexts of use of the expressions."*[9]
>
> — Montague, 1968

5 Although there are many expressions that do not contribute to the truth conditions of sentences but are still imbued with "meaning" in the sense that they are not fully (or satisfactorily) substitutable to truth conditionally equivalent linguistic units. Think e. g.of the distinction between "and" and "but".

6 "La linguistique a pour unique et véritable objet la langue envisagée en elle-même et pour elle-même.", though to be fair, Saussure and most structuralists initially avoided semantics and lexicology, as underlined in and until Hjelmslev (1957)

7 "Entité autonome de dépendances internes" Hjelmslev (1957, p.100)

8 "Toute description scientifique présuppose que l'objet de la description soit conçu comme une structure" Hjelmslev (1957, p.101)

9 The same distinction is made in Morris (1938), with the exception that the approach presented here focuses exclusively on linguistic signs and not signs in general.

Importantly, however, truth conditions do not give a *truth value*. Even the most fervent supporters of this approach know very well that purely referential semantic meaning can only ever be computed if we take into account some degree of context. *Indexicals*, for example, are a class of terms whose reference depends on the context of utterance. Words like *I, you, here, now*, among others, are indexicals; if the truth conditions of a sentence are to be computed, their reference has to be defined, and it can only be defined in case the listener has information that goes beyond the linguistic signal itself.

This leads us to a first issue when distinguishing semantics from pragmatics: there is a common approach which states that the former focuses on situation independent content et the latter on situation dependent content. This would lead to a classification of the study of indexicals in pragmatics and not semantics. This approach can be for example found in a first reading of Szabo (2006):

> *"Semantics is the study of meaning, or more precisely, the study of the relation between linguistic expressions and their meanings. [. . .] Pragmatics is the study of context, or more precisely, a study of the way context can influence our understanding of linguistic utterances."*
>
> — Szabo, 2006

But in Gazdar (1979)[10] , we find the concise and to the point:

> *"PRAGMATICS = MEANING — TRUTH CONDITIONS"*
>
> — Gazdar, 1979

Now this would entail that the study of indexicals is a semantics matter, since they contribute to truth conditions. Gazdar (1979) is not the only work that goes for this distinction instead of the situational/non-situational one. For example, in Korta and Perry (2020), we also find the following:

> *"Semantic information is information encoded in what is uttered — these are stable linguistic features of the sentence — together with any extralinguistic information that provides (semantic) values to context-sensitive expressions in what is uttered. Pragmatic information is (extralinguistic) information that arises from an actual act of utterance, and is relevant to* the hearer's determination of what the speaker is communicating. *Whereas semantic information is encoded in what is uttered, pragmatic information is generated by, or at least made relevant by, the act of uttering it."*[11]
>
> — Bach, 2008

This last position underlines a key point of pragmatics that we will come back to soon, it also makes use of the term *act*, which will be the focus of the very next part.

10 Quoted in Korta and Perry (2020).
11 Emphasis mine.

In fact we can go further than indexicals. Sentences as simple as (6) already imply some notion of time and space that is not found in the signal itself. The computation of a truth value will rely on that *unarticulated* content (Perry and Blackburn, 1986).

(6) It is raining.

The idea here is that it can be argued that the truth conditions of a sentence are always to some extent dependent on contextual information. In fact, once we do pragmatics, we are no longer interested in sentences like this one, but in *utterances*. Sentences are then seen as utterance types, purely grammatical and theoretic objects abstracted from any non-linguistic elements. If one agrees to the idea that such an object can hardly have clearly defined truth conditions, then sentences conceived as such are closer to a syntactic object than anything else. The *meaning* of a sentence then becomes the set of truth conditions that can indeed be extracted from the pure linguistic form, but it becomes difficult to say that a sentence *has meaning*. We might rather say that it has *extension*.

> *"A meaning for a sentence determines the conditions under which the sentence is true or false. It determines the truth-value of the sentence in various possible states of affairs, at various times, at various places, for various speakers, and so on. [...] We call the truth-value of a sentence the extension of that sentence; we call the thing named by a name the extension of that name; we call the set of things to which a common noun applies the extension of that common noun. The extension of something in one of these three categories depends on its meaning and, in general, on other things as well: on facts about the world, on the time of utterance, on the place of the utterance, on the speaker, on the surrounding discourse, etc. It is the meaning that which determines how the extension depends on the combination of other relevant factors."*
>
> — Lewis, 1970

Adopting that point of view, the role of semantics is still the computation of truth conditions, but a sentence is generally thought to have no truth value, only an utterance does, because the truth value of the sentence is dependent on a number of variables, many of which are contextual.

So what we generally understand as *meaning* is not exactly what is studied in semantics. Semantics gives us *possible meanings*, but because it focuses on sentences, which are linguistic objects abstract from world usage, it does not tell us what has been said, right now, by that person. This would be the role of pragmatics.

By way of conclusion on these remarks: the distinction between semantics and pragmatics is usually made through the prism of situation-dependence, and semantics has indeed focused on the referential meaning of expressions and the equivalence of *meaning* and *truth conditions*. This is due, on the side of linguistics, to the desire of envisioning language meaning as an autonomous analyzable object. On the side of philosophy of language and logic, this is due to the tradition in the semantics of formal languages which it was hoped would be applicable

to the semantics of natural languages. The distinction between what constitutes context information and what does not is not as clear-cut as we would initially think, however, and items like indexicals and others already point towards a relative permeability of the two realms.

> *"I said that semantics is the study of meaning and pragmatics the study of context. This makes it seem as if they are about entirely different things. In a way, this is so: primarily expression types have meaning and expression tokens occur in contexts. Courtesies are extended in both directions, but tokens can only be said to have a certain meaning by extension, in virtue of being tokens of a type with that meaning, and types can only be said to occur in a context by extension, in virtue of being types to which a token that occurs in that context belongs. Despite their differences there is a way to pull meaning and context together: they are the two sources of information used in interpreting utterances. An utterance is an* action *involving the articulation of a linguistic expression* by an intentional agent*, the speaker, directed* at an intentional agent*, the addressee. The interpretation of the utterance is a certain cognitive process whereby the addressee ascertains what the speaker meant in making the utterance."*[12]
>
> — Szabo, 2006

It feels more fruitful, then, to define pragmatics not simply as *meaning in context* or *the study of context*, but as *the study of utterances* and of *intention*, if the distinction is to be made with semantics[13]. In any case the two approaches are complementary when studying the meaning of linguistic signals.

Despite this, one key motivation in the emergence of pragmatics as a field was not to be found in considerations about indexicals and context-sensitivity of truth conditions (what is called *near-side* pragmatics in Korta and Perry, 2020), but in a much more radical approach distinguishing many flavors of *what is said*.

2.2 COMMUNICATING INTENTION

2.2.1 Speech acts

The academic study of pragmatics as a field of linguistics are generally said to start with Austin (1962) and Searle (1969), it comes from the opposition to an analysis that sees many (most) utterances as statements and descriptions, and focuses notably on considerations about *performative utterances*.

(7) I name this ship the *Queen Elizabeth* (Austin, 1962)

If (7) is uttered in the right circumstances (i. e.smashing a bottle against the stem of an unbaptized ship), then even though syntactically it has all the features of a descriptive sentence,

12 Emphases mine.

13 For arguments against this distinction, see e. g.Lepore and Stone (2015) and Parikh (2010).

it can hardly be said that it's a description. The speaker here is not *describing* any event, they are accomplishing the naming of the ship with the utterance itself. This is what a "performative sentence" (Austin, 1962) is, an action performed by the utterance of the sentence itself. The initial distinction in utterances made by Austin is that between what he calls *constatives* – more or less all sentences that look like statements, even when they aren't – and *performatives*, but the main point in his initial approach is to underline the fact that using language is not just about describing the world. Even more than this, actually: using language is not necessarily about using language, but about accomplishing actions.

Following Austin's opening William James lecture, *speech act theory* gradually emerged. The goal here was to propose another sense to the notion of linguistic meaning. Sure, linguistic signals have a semantics, they have content, but their *meaning* is to be found in social interactions. Language is used to do things, things that are beyond the mere communication of ideas and concepts. It is used to influence the state of the world and the actions of the agents within it. The approach put forward in Austin (1962) operates a shift in the way we think about language and what it means to use it. Suddenly, notions of *truth value* are no longer relevant, the notion of *felicitousness* becomes more appropriate: whenever the intention of the speaker in producing an utterance is recognized by the addressee(s), then the speech act is felicitous.

More precisely, Austin divides speech acts into three distinct parts:

1. *Locutionary acts*: the act of producing the linguistic signal itself.

2. *Illocutionary acts*: the intention behind the production of the locutionary act (e. g.an order, a warning, etc.).

3. *Perlocutionary acts*: the effects of the locutionary act on the addressee (e. g.obedience, fear, etc.)

In Searle (1969), speech act theory is improved upon and a vision of pragmatics as the study of what is conveyed by language *beyond* language emerges. Its focus is mostly on the conditions that ensure the felicitousness and success of illocutionary acts, that is, under which conditions do they emerge/are understood by the addressee. More generally, the view of linguistic meaning adopted by speech act theory revolves around social interactions and the social use of language. The notion of *intention* is present, but it is only the *intention* to act in a certain way that is at play here.

Although we do not generally adopt a speech act-centered approach to pragmatics in this work, it is important to note two insights brought to us by speech act theory that will be important here:

- There is meaning beyond the content of the message. In describing language use in a given community, it might be fruitful to focus not on *what is said* – in the sense of "what is the literal signification of that utterance?" – but on *what is done* when producing a particular utterance.

- The form of the linguistic signal itself is not necessarily enough of an indicator to convey *what is done*. This point is key in understanding what dogwhistles are and how they function, as we will argue in part iii. Dogwhistles as an act of language are to be seen under the scope of duplicity, and the illocutionary act that is in fact performed by the speaker is not necessarily in keeping with what is understood by the listener.

Importantly, one of the original intents in Austin (1962) was to evade the conception of language as an autonomous structure. That being said, a lot of the theory subsequent to his work has focused on felicity conditions of given speech acts, much to the detriment of the actual study of social interactions among users of language. Illocutionary acts have over time been treated, in some way, as autonomous entities, ignoring the realities of social interactions and not acknowledging the fact that whether an illocutionary act was successful is hugely dependant on the structure of the social world in which it is attempted. In most cases, not all speakers have the same chances to utter some of these acts. *Identity and status play a role in message interpretation* (Pratt, 1986)[14].

Speech act theory was a turning point in the study of pragmatics and linguistic meaning in general, shifting the focus away from message itself and focusing on its *intention* and *reception*. These two concepts are key in a modern understanding of pragmatics. Parallel to the social, interaction-oriented approach promoted by Austin (1962) and Searle (1969) came a psychological, individual-centered approach to pragmatics focusing on this notion of intention and how it has in fact an impact also on *what is said*, bringing the study of pragmatics back to more language-centric considerations.

2.2.2 Implicatures *and ambiguity resolution*

Some time after speech act theory emerged, Grice (1967) also discussed meaning beyond the linguistic signal, but did so in a fairly different way, focusing notably on *propositional meaning* inferred from linguistic input.

Grice is mostly known for the Cooperative Principle (CP) and the *maxims of conversation*, respectively a principle and a set of four rules to be followed to ensure fruitful conversation and information communication. A large part of contemporary pragmatics is rooted in Gricean analyses, and in many ways so is this work. All of this stems from the concept of *implicature*.

> *"Our talk exchanges do not normally consist of a succession of disconnected remarks, and would not be rational if they did. They are characteristically, to some degree at least, cooperative efforts; and each participant recognizes in them, to some extent, a common purpose or set of purposes. [...]* Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged."[15]

14 The idea that social identity plays a part in the interpretation of linguistic signals is one of the ideas behind the notion of *social meaning*, which will be discussed in Chapter 3.

15 Emphasis mine. The emphasized sentence is the common formulation of the CP.

— Grice, 1967

The CP thus results from the intuition that successions of utterances in a conversation are connected to each other even though it might not appear to be the case. In other words, listeners infer from the linguistic signal information that is not there in the first place. This supplementary information is what is called "implicature" in Grice (1967).

A first kind of implicatures that is hinted at in Grice (1967) are *conventional implicatures*, which rely on the *conventional meaning* of words and phrases. One way of seeing those is to compare sentences with identical truth conditions that nonetheless seem to differ in meaning. The analysis found in Grice (1967) is that this difference is due to convention[16].

One way of seeing those is to consider words that convey attitudes in addition to having content. One can think of swear words in general, like "fuck", which can convey e. g.frustration, but many of these do not necessarily have *content* per se. A clearer example of what we can think of when invoking the concept of *conventional implicature* is that of slurs. Slurs are a way to refer to individuals through the community to which they belong (and therefore have an extension) in a derogatory way (conveying an attitude towards that group). A Gricean reading of the interpretation of slurs would be that the "derogatory" part of the meaning is a conventional implicature. So for example, a sentence like (8) would trigger the two interpretations in (9):

(8) Fargoth is such a n'wah[17]

(9) a. Fargoth is a foreigner.

b. The speaker dislikes foreigners.

In this context, "n'wah" and "foreigner" have the same extension, but there is an attitude conveyed by "n'wah" towards foreigners that is not conveyed by "foreigner". This attitude can be anayzed as being part of the conventional meaning of "n'wah" and any proposition regarding the attitude of the speaker regarding foreigners from the use of "n'wah" would be a conventional implicature[18].

Importantly in the context of this work, going against a first analysis presented in Stanley (2015) and following remarks found in Henderson and McCready (2018), we do not consider

16 As in footnote 5, you can think of the distinction between "and" and "but" here. A Gricean reading of this would be that the meaning of *concession* carried by "but" is conventional, and any proposition concerning this concessiveness made from the presence of "but" in the sentence would be a conventional implicature. This is one reason why conventional implicatures as a concept are sometimes decried in pragmatics. Surely that concessive meaning *is* part of *what is said* and not merely an implicature. Nonetheless, there are other examples that are more problematic (like the slur example following this note). See Bach (1999) for arguments against the existence of conventional implicatures as a pragmatic object.

17 The term *n'wah* is a slur used by the natives of the island of Vvardenfell of dunmer descent to refer to foreigners in the Elder Scrolls video game series. According to a variey of sources, it can mean "foreigner" or "slave" and is generally used as an insult. It was chosen with the objective of not offending anyone reading this.

18 This is obviously a very simplified picture of how slurs can be thought to work, but in general the phenomenon, along with swear words more generally, is used as a way to illustrate the need for multiple levels of meaning, see for example Gutzmann and Gärtner (2013), McCready (2010), and Potts (2007).

dogwhistles to trigger conventional implicatures, and the reason for that is that conventional implicatures are generally analyzed as being *non-cancellable*. We will discuss this more in part iii.

The focus in Grice (1967) is rather on *conversational implicatures*, which are implicatures not triggered by the conventional meanings of words but by the act of utterance itself. One key insight that is found in the Gricean approach is that if you assume that people follow some version of the CP and the maxims of conversation, then the conversational implicatures can be computed. This finally brings us to the maxims, of which there are four[19]:

- Quantity:
    1. Make your contribution as informative as is required (for the current purposes of the exchange).
    2. Do not make your contribution more informative than is required.
- Quality:
    1. Do not say what you believe to be false.
    2. Do not say that for which you lack adequate evidence.
- Relation: Be relevant.
- Manner:
    1. Avoid obscurity of expression.
    2. Avoid ambiguity.
    3. Be brief (avoid unnecessary prolixity).
    4. Be orderly.

It is unclear what the maxims actually are, whether they are normative or descriptive, whether they are all necessary, etc. (Korta and Perry, 2020), nonetheless, they serve as a starting point for the inference of propositions that go beyond *what is said*. It is not said in Grice (1967) that the list is definitive or comprehensive; in some way, the important point made by the paper is not the maxims themselves, but is rather as follows:

- Not everything that is understood by listeners is actually said during a conversation.

→ Listeners draw conclusions (*implicatures*) based on what is said to them, in addition to understanding the utterances themselves.

→ There is regularity in these implicatures. Their computation seems to follow a logical pattern.

19 These are the original formulations of Grice (1967).

→ There are ways for us to describe the first principles from which that logic allows the computation of implicatures.

→ The (possibly unconscious) awareness of these first principles, in turn, determine the utterances that are produced.

Conversational implicatures are implicated (by the speakers) and inferred (by the listeners) following a set of rules, which allows each competent user of language to *mean* more than they *say*. In the Gricean approach, any form of sentence can carry any kind of implicature given the right context. For example, sentence (10) will trigger different conversational implicatures depending on whether it is answering question (11a) or (11b)[20].

(10) The neighbors were having a party.

(11) a. Slept well last night?
(10) $\rightsquigarrow$ No, I did not sleep well last night.

b. Did anything last night?
(10) $\rightsquigarrow$ I went to my neighbors' party.

Contrary to conventional implicatures, conversational implicatures can be easily canceled without causing any problem. You could very well imagine, for example, that (10), uttered as a response to (a), would be followed by "but oddly enough, I did sleep well".

There is a category of conversational implicatures that is almost systematic and always triggered by the same linguistic forms. Grice (1967) calls them Generalized Conversational Implicature (GCI). The idea is the same as regular conversational implicatures, except that they seem to be much more specific. One such GCI that will be of particular interest to us in this work, notably because it has spawned a host of theories and models, are so-called *scalar implicatures*.

The idea behind scalar implicatures is as follows: if a speaker S utters a sentence p such that there exists a sentence p' that asymmetrically entails p, then S implicates $\neg p'$. We say a proposition *entails* another one when that second proposition is necessarily true given the first one. *Asymmetrical entailment* of p by p' means that p' entails p but that p does not entail p'. Let's illustrate with an example.

(12) a. I ate some of the apples.

b. I ate all of the apples.

In that case, (b) asymmetrically entails (a), because it is always the case that if I ate all of the apples, then I ate some of them, but it is not always the case that if I ate some of the apples, then I ate all of them. A scalar implicature is when saying (a) actually implies that (b)

20 The symbol $\rightsquigarrow$ is used to mean "implicates".

is not true. The Gricean reading of scalar implicatures is that they are related to the maxims of quantity and quality: there exists an alternative quantifier that gives me a more precise information about the situation; that more precise quantifier has not been used; I conclude that the supplementary information it brings must be false (or unknown), because if it were information believed to be true, then in virtue of the maxims of quality and quantity, that more precise quantifier would have been used.

This is, again, the simpler picture. Since Grice (1967), there has been work underlining, notably, that we can have many scales of asymmetrical entailment even for a word as simple as "some". For example:

(13) I ate several of the apples.

The word "several" asymmetrically entails "some" in the same way that "all" does, and yet it does not appear that the negation of "several" is ever implicated when using "some". This is called the "symmetry" problem in Lepore and Stone (2015).

In fact, there are several such limitations to the original Gricean programme taken as such , and it has been acknowledged for a time now that it is insufficient to give a clear picture of what is going on with implicatures. The symmetry problem just mentioned is an important shortcoming of the theory, but it is not the only one presented in Lepore and Stone (2015), which takes a stance against the very notion of conversational implicatures.

Two main approaches to pragmatics have emerged upon the discovery of these limitations. The first are so-called neo-Gricean accounts, which keep as a starting point some idea of maxim to be followed. The second approach dismisses maxims focused on language use and rather put forward the idea that implicatures and general meaning inference are related to broader cognitive abilities (Sperber and Wilson, 1995; Wilson and Sperber, 2002). This work does not bring an answer regarding which approach is more fruitful, but the formal models that we will focus on and the general tradition in game-theoretic pragmatics that we are interested in are closer to neo-Gricean approaches than they are to i. e.relevance-theoretic approaches. Notably, models such as RSA, to be presented shortly, have initially focused on scalar implicatures and explanations of the phenomenon as it is presented in Gricean and neo-Gricean approaches.

No matter the philosophical stance taken here as a starting point, the models that we have used to describe dogwhistles are inspired by models initially used to describe scalar implicatures, which itself is one phenomenon related to the notion of *ambiguity resolution*, a name given to processes through which the reference of an expression is made more precise than it initially is by eliminating improbable interpretations[21].

21 The term *ambiguity* is here used a bit loosely. In fact, we might rather want to say that the interpretation of scalar implicatures results from a case of *vagueness* or *underspecification* more than actual *ambiguity*. I think it can be argued that these phenomena are generally close, and that in any case formal models like RSA can be thought to be explanatory for all of these phenomena, which rely on similar underlying interpretation processes to be resolved (when they are). For a clear and concise illustration of the difference between these things, see Sennet (2021, section 2).

Hopefully this seciton has cleared a little bit what the focus of pragmatics is/can be, in more precise terms than what was discussed earlier. The key notions that will be useful to us in the rest of the dissertation are those of *implicature* – and specifically *scalar implicature* – of *speaker intention*, of *ambiguity resolution* and of *speech act*. One especially important point that has to be underlined too is that traditionally, the study of pragmatics in the line of Grice has focused on a notion of *cooperation*, but we can invoke other principles besides a principle of cooperation to account for how communication between agents unfolds. In Leech (1983), we find mention of other principles that can explain how certain linguistic forms are favored over others, principles like *politeness, clarity, economy* and others, which can all serve to explain speech behavior.

In fact, it will be safe to say that a notion like *cooperation*, at least in its more common sense, is not necessarily appropriate to discuss dogwhistles. Nevertheless, the CP and Gricean pragmatics have been among the most influential approaches to pragmatics, and the notion of cooperation at least in the sense of converging on an interpretation will still be an interesting starting point for the analysis of many phenomena.

2.3 FROM LANGUAGE GAMES TO GAME THEORY

That last sentence illustrates one concept of dialogue that has been seminal in the elaboration of game-theoretic pragmatics, the current of thought that this work is closest to.

The idea of language being envisioned as a *game* is usually associated with Wittgenstein (1953). In this work, Wittgenstein's approach to language is reminiscent of what we have seen with speech acts in that language is envisioned as a medium to accomplish actions in a given context. What Wittgenstein calls "language games" are in a sense subsets of the human experience where language will play a part. In those games, language is used by a player to trigger a behavior in another player, but the *language* acquires its meaning and importance through the game. This means that any given message can have a variety of uses that will be defined by the context in which it is uttered.

This places Wittgenstein's considerations close to Austinian pragmatics: the context gives the meaning. This is the reasoning behind the famous:

> *"If a lion could talk, we could not understand him."*
>
> — Wittgenstein, 1953

Lions do not have any shared experiences with us, they do not take part in the same *games* as we do, therefore even if they had knowledge of a human language, communication with them would be impossible.

Wittgenstein (1953) happened to use the term *game*, but it is rather to Lewis (1969) that we owe the contemporary approach to language as a *game*; specifically as a *signaling game*.

In Lewis (1969), *convention*, taken in a fairly general sense, is described in game-theoretic terms. More specifically, it is described as a "coordination game". Key aspects of language[22], like the determination of a specific grammar, the semantics of natural language, etc. are subsequently seen also as coordination games. The games used by Lewis were later called "Lewis signaling games".

2.3.1 *Communication as a signaling game*

Much more so than what is presented in Wittgenstein (1953), Lewis (1969) presents us with a formalization of the concept of game when applied to communication in game-theoretic terms[23]. The concept of a signaling game has thus been formalized using the vocabulary and objects found in Game Theory (GT). This first impulse has spawned a rich literature on the subject (Jäger, 2008; Parikh, 2001; van Rooij, 2004) and it is now relatively standard to analyze interactions in terms of games.

A signaling game is a game with two players: a *sender* and a *receiver*. The sender is aware of some state of the world that the receiver ignores. In the present case, we are considering signaling games for coordination, which in some sense are related to *cooperative* game theory, a name that echoes the CP that we have mentioned earlier. In the coordination version of signaling games, the winning state is achieved for both players when the receiver manages to correctly retrieve the state of the world from the action that the sender chooses to do.

Formally, as per Franke (2009), a signaling game with meaningful signals is a tuple:

Definition 2.3.1. Signaling game

A signaling game with meaningful signals is a tuple of the form:

$$\langle \{S, R\}, T, Pr, M, [\![\cdot]\!], A, U_S, U_R \rangle$$

In this tuple, we find the following:

- S and R are the *sender* and the *receiver*, the two players of the game.
- T is a set of states of the world.
- $Pr \in \Delta(T)$ is a full-support probability distribution over T, representing the receiver's uncertainty about which state in T is actual.
- M is a set of messages available to the sender.

22 And communication in general.

23 This section and subsequent sections will rely on some knowledge of game theory. While it is not necessary to be an expert, it will be useful to know some of the basic notions and key concepts behind it. See Appendix A for a very concise introduction to some concepts and see Hargreaves-Heap and Varoufakis (2004), Osborne (2004), Ross (2019), and Tadelis (2013) for more in-depth introductions both to the philosophy behind it and to the mathematical theory itself.

$[\![\cdot]\!]$	$\exists\neg\forall$	$\forall$
$[\![\text{some}]\!]$	1	1
$[\![\text{all}]\!]$	0	1

Table 1: The interpretation function for the some-all apples game presented in section 2.3. A cell containing a 1 is to be read as "the message m is this row comprises the state of the world t of that column in its extension" or $t_{col} \in [\![m_{row}]\!]$. A 0 means $t_{col} \notin [\![m_{row}]\!]$.

- $[\![\cdot]\!] : M \to \mathcal{P}(T) \setminus \emptyset$ is a *denotation function* or *interpretation function*. It gives the semantic meaning of a message m, where semantic meaning is understood as the extension of m. Meaning it gives all the states of the world t that are compatible with m.
- A is the set of response actions available to the receiver.
- $U_{S,R} : T \times M \times A \to \mathbb{R}$ are utility functions for the sender and the receiver that give a numerical value to each possible outcome of the game. The ordering of those utility values is supposed to represent the preferences of each player.

This configuration is the basis for a lot of work in game-theoretic pragmatics, including RSA and SMG models. A standard approach to how this applies to linguistics is to consider the case of scalar implicatures, more specifically the case presented in (12), reproduced (and slightly augmented) here:

(14) a. I ate some of the apples.
$\rightsquigarrow$ I did not eat all of the apples.

b. I ate all of the apples.

The classical analysis in terms of implicatures is that (a) triggers the "not-all" implicature because it is assymmetrically entailed by (b). How can we translate that into our signaling game framework? Franke (2009) and Scontras, Tessler, and Franke (2018) provide very clear and thorough presentations. I will not attempt to outdo those resources here but will still provide an illustration (hugely influenced by them).

Let $M = \{\text{all}, \text{some}\}$ be our set of possible messages, each corresponding to the appropriate version of (14). Let $T = \{\exists\neg\forall, \forall\}$ be the set of possible states of the world. This translates to "some not all" and "all".

The receiver knows nothing of the world and $\Pr$ is therefore uniform over T.

In the case presented here, the actions that the receiver can do are *interpretations*. Right now, we will interpret those as a choice that the receiver makes in T.

Now that we have set all of these, $[\![\cdot]\!]$ can be found in Table 1.

Now what we need is to represent the possible outcomes. We give those in extended form and normal form in, respectively, Figure 1 and Table 2. Note that in general, which message is used by the sender is irrelevant in the winning of the game itself. As long as the action made by the receiver is consistent with the state of the world, then the game is won, no matter

OUTCOMES	$a_{\exists\neg\forall}$	$a_{\forall}$
$t_{\exists\neg\forall}$	1, 1	0, 0
$t_{\forall}$	0, 0	1, 1

Table 2: The $T \times A$ outcomes and their utilities.

which message was sent. This is why in Table 2, we do not represent the messages themselves. The goal of the message is to influence the behavior of the receiver, but it is ultimately the behavior of the receiver which decides the final outcome.

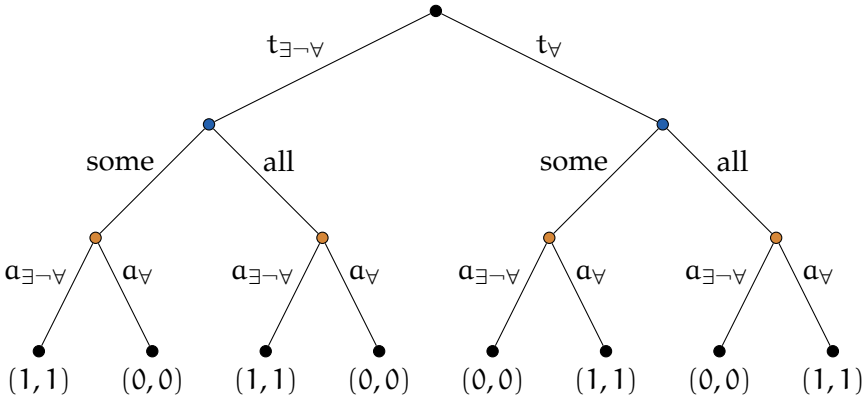


Figure 1: Extended form representation of the some-all game with apples from section 2.3.

We can see from these representations that there are several possible equilibria to such a game. Notably, having words be systematically interpreted to mean the opposite of what they mean can be an equilibrium. For an overview of all the many equilibria of this game (Nash and Pefect Bayesian), see Franke (2009), where an in-depth analysis of this is conducted. Whether we apply Perfect Bayesian equilibrium or modified versions where some notion of cost and respecting the maxim of quality are respected, we still end up with some issues and unwanted equilibria.

What we need is a method that, taking into account the fact that we have the message set that we have, allows the receiver to make the correct interpretation nonetheless upon hearing the message. In other words, we need a model that tells us how to systematically reach the situation where upon hearing "some", the action $a_{\exists\neg\forall}$ is taken.

2.3.2 *IBR and RSA*

This is the ambition behind the IBR and the RSA models. Both models rely on the same mechanisms, that can be summed up as follows: both senders and receivers have prior beliefs about the behaviors of each other. They also have knowledge of this. Recursive thinking about this allows for disambiguation and emergence of scalar implicatures. Because they are fairly similar and the latter is more central to our approach, this is the one we will focus on here.

As stated in Scontras, Tessler, and Franke (2018), the point behind RSA models in particular is to reverse the tendency historically observed in the study of linguistic meaning that sought, through compositionality, to reach the semantic meaning of utterances, seen as the finality of meaning computation. In the tradition of formal pragmatics that RSA follows, semantics is recognized "not as one of the final steps in meaning calculation, but as one of the first."

How does this work? The basic idea is rather simple. Each player has an internal representation of the other player. At the beginning of the process, a *literal listener*[24] L_0 is envisioned. This literal listener uses a set of prior beliefs about the world and otherwise treats each signal in a literal manner.

This literal listener is the listener envisioned by the *pragmatic speaker* S_1. The pragmatic speaker, knowing how a literal listener would interpret their messages, will act accordingly.

Then the *pragmatic listener* L_1 is a listener who assumes that the speaker in front of them is a *pragmatic speaker*. This can go on recursively for as long as one wants, but typically is stopped at this step. We therefore have at the same time a *speaker model*, giving us the behavior of a speaker, and a *listener model*, giving us the behavior of a listener. Importantly, the *literal listener* is thought to be a theoretical object of sorts. It is a conceptual object used by both speaker and listener to use signals in a certain way/correctly infer the state of the world from those signals. The literal listener is who a pragmatic speaker *believes* they are talking to. Meaning that a S_1 speaker does not believe that the listener in front of them will exhibit scalar implicature behavior, yet even then, the winning strategy for them will be to use the messages at their disposal in an unambiguous way. The message "some", already with S_1, is used as if $[\![\text{some}]\!] = \{t_{\exists\neg\forall}\}$ and not $\{t_{\exists\neg\forall}, t_{\forall}\}$.

This will all be clearer with an example. Frank and Goodman (2012) and Scontras, Tessler, and Franke (2018) start with the example of a reference game[25], but we can jump straight into scalar implicature territory here (Goodman and Stuhlmüller, 2013).

Let Γ be a tuple: $\langle \{S_1, L_1\}, L_0, T, Pr, M, [\![\cdot]\!], A, C, U_S, \alpha \rangle$.

- S_1 is a pragmatic speaker and L_1 is a pragmatic listener.
- $L_0 : M \to \Delta(T)$ is a function from messages to probability distributions over states of the world. This distribution is determined by the interpretation function $[\![\cdot]\!]$ and prior beliefs Pr, in the following manner: $L_0(t|m) = [\![m]\!](t) * Pr(s)$.
- $T = \{1, 2, 3\}$ is a set of possible states of the world.
- $Pr \in \Delta(T)$ is the a priori beliefs over possible worlds. We set it to be uniform.
- $M = \{\text{some, all}\}$ is the set of possible messages available to the speaker.

24 The terminology used in the RSA literature is that of *speaker* and *listener* rather than the traditional *sender* and *receiver* of signaling games. Starting from now, we will use the RSA terminology as well.

25 A game where the listener will have to find the object that a signal is referring to with a speaker using possibly ambiguous words. These games constitute many of the examples of what Wittgenstein (1953) calls "language games".

$[\![\cdot]\!]$	1	2	3
$[\![\text{some}]\!]$	1	1	1
$[\![\text{all}]\!]$	0	0	1

Table 3: The interpretation function for the apples game analyzed through the lens of the RSA framework.

$L_0(t\|m)$	$L_0(1\|m)$	$L_0(2\|m)$	$L_0(3\|m)$
$L_0(t\|\text{some})$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
$L_0(t\|\text{all})$	0	0	1

Table 4: The literal listener for the apples game analyzed through the lens of the RSA framework.

- $[\![\cdot]\!] : M \times T \to \{0, 1\}$ is an interpretation function[26].
- $A = \{1, 2, 3\}$ is the set of interpretations available to the listener.
- $C : M \to \mathbb{R}$ is a *cost* function. It assigns a cost to every message, representing some measure of impracticality of the message in question. In this work, we will mostly assume that messages are *costless* and this function will not be used.
- $U_S : M \times T \to \mathbb{R}$ is the speaker's utility function, defined as $U_S(m; t) = \log(L_0(t|m)) - C(m)$[27]
- $\alpha \in \mathbb{R}$ is a *temperature* parameter. It is used to quantify the optimality of a speaker in choosing their messages, the higher the alpha, the more optimal the speaker will be. In this example, we will set $\alpha = 1$.

$[\![\cdot]\!]$ is defined in Table 3 and L_0 in Table 4. This is what we start with, a set of messages with a precise extension in terms of possible worlds, and a literal listener that takes the literal interpretation of each of those messages (and is therefore unable to tell which of the three possible states is more likely to be true in case they hear "some").

That literal listener can be thought of as a listener model for the pragmatic speaker S_1, i. e.how they assume the listener will act[28]. Given this, the pragmatic speaker will choose which message to send using the utility U_S associated with each message. These utility scores

26 This is the Boolean semantics description of the interpretation function. Given a message m and a state of the world t, this gives out 1 if t is part of the extension of m, 0 otherwise. This is equivalent to what we described earlier as an interpretation function but has the added benefit of giving us directly 1 or 0.

27 This is based on *negative surprisal*, a measure defined in information theory. It is standard to use it in such frameworks as RSA. If $L_0(t|m)$ is at 1 (one possible interpretation of the message m), then $U_S(m; t) = 0$, and as $L_0(t|m)$ approaches 0, $U_S(m; t)$ approaches $-\infty$. The idea here is that speakers will try to avoid using messages that only *sometimes* describe t and will favor messages that *often* describe t to avoid surprising listeners. Highly surprising messages here have extremely low utility.

28 Note that in this case, and as underlined in Franke (2009) when talking about IBR's S_1 and R_1, this model of a speaker will act pragmatically but will not have the belief that the listener in front of them does. If we were to apply this to the real world, choosing at least S_2 to be a model for the speaker would philosophically be more motivated.

$P_{S_1}(m\|t)$	$P_{S_1}(m\|1)$	$P_{S_1}(m\|2)$	$P_{S_1}(m\|3)$
$P_{S_1}(\text{some}\|t)$	1	1	0.25
$P_{S_1}(\text{all}\|t)$	0	0	0.75

Table 5: The pragmatic speaker for the apples game analyzed through the lens of the RSA framework.

are then interpreted as log-probabilities and we can derive the actual probability of sending a message rather than another using the *soft-max choice rule*[29]:

$$P_{S_1}(m|t) \propto exp(\alpha * U_S(m;t))$$

When fully expanded:

$$\frac{P_{S_1}(m|t) = exp(\alpha * (log(L_0(t|m_i)) - C(m_i)))}{\sum_j exp(\alpha * (logL_0(t|m_j) - C(m_j)))}$$

For all possible t (remember: the speaker has observed t), this gives us the results in Table 5. We can see already that assuming that they are facing a literal listener, the pragmatic speaker will already act in a way that is reminiscent of scalar implicatures, avoiding the use of "some" when $t = 3$.

The pragmatic listener L_1 computes the probability of $t = x$ given some message m. This is when we switch to a listener model, where this time the pragmatic speaker is the speaker model for the listener, their belief about how the speaker is going to act. They also take into account their prior beliefs about the state of the world to do so, leaving us with:

$$P_{L_1}(t|m) \propto P_{S_1}(m|t) * Pr(t)$$

Fully expanded:

$$P_{L_1}(t|m) = \frac{P_{S_1}(m|t_i) * Pr(t_i)}{\sum_j P_{S_1}(m|t_j) * Pr(t_j)}$$

The interpretations of L_1 are presented in Table 6. We can see that it is now very unlikely that "some" will be interpreted as referring to $t = 3$. We have derived a scalar implicature behavior.

SOME REMARKS The model sketched here (as well as the IBR model) has the advantage of always converging to the desirable equilibrium in the some-all game (Franke, 2009; Scontras, Tessler, and Franke, 2018), it also gives us a potential mechanism behind the reaching of

29 Why do we use soft-max and not a more regular normalization process? This is relatively standard in the modeling of decision-making. One reason is that we have chosen to interpret utilities as log-probabilities. This means that in order to have regular probabilities it is necessary to go through the exponential function. Another reason is that it gives us more flexibility. Normalization only takes into account proportions, such that multiplying our utilities by our temperature parameter α will have no effect on the final result, whereas the soft-max function also takes into account the actual values of the utility, and underlines differences in actual utility scores. While a regular normalization would treat the scores $(1, 2)$ and $(10, 20)$ the same when converting them to probabilities as $(0.3\bar{3}, 0.6\bar{6})$, the soft-max function would not, converting the first pair into $\approx (0.28, 0.73)$ and the second into $\approx (0.01, 0.99)$.

$P_{L_1}(t\|m)$	$P_{L_1}(1\|m)$	$P_{L_1}(2\|m)$	$P_{L_1}(3\|m)$
$P_{L_1}(t\|some)$	$0.4\bar{4}$	$0.4\bar{4}$	$0.1\bar{1}$
$P_{L_1}(t\|all)$	0	0	1

Table 6: The pragmatic listener for the apples game analyzed through the lens of the RSA framework.

that equilibrium. That being said, the present illustration makes a lot of strong assumptions, including: the shared knowledge that there are only 3 apples, the limitation of the vocabulary to 2 words, etc.

Some of these are addressed in Scontras, Tessler, and Franke (2018), where more intricate situations are presented. For the purposes of this work, however, this short presentation will suffice to understand the models that we will propose in iii.

2.4 CONCLUDING REMARKS

In this part, we have[30] fairly concisely seen what pragmatics is, given some informal definitions of many of its core concepts, seen what scalar implicatures are, and some examples of attempts at formalizing those concepts. This is not supposed to be a substitute to an actual introduction to all of these topics, but it might help uninitiated readers to understand a bit more what we mean by *game-theoretic pragmatics*, while also making some of the objects that will be discussed later in this work grounded in earlier literature.

If the reader is interested in knowing more, the many references of this part can be useful, Franke (2009) provides a nice introduction, and Jäger (2008) gives a nice overview of the applications of game theory to questions in linguistics (including the use of evolutionary game theory to treat issues of language emergence, which we have not discussed here).

The next chapter will focus on fairly different matters, as we will discuss concepts that have historically emerged in a related, yet very different, field. The theoretical and technical tools presented here will however be central when discussing our analysis of dogwhistles, and will be necessary even to the proper understanding of SMG, an attempt to use those same tools on very different conceptual objects.

30 I hope.

3

SOCIAL MEANING

"And when any Ephraimite who escaped said, "Let me cross over," the men of Gilead would say to him, "Are you an Ephraimite?" If he said, "No," then they would say to him, "Then say, 'Shibboleth'!" And he would say, "Sibboleth," for he could not pronounce it right. Then they would take him and kill him at the fords of the Jordan. There fell at that time forty-two thousand Ephraimites."

— Judges 12:5-6, illlustrating how pronunciation can index identity (*Holy Bible: The New King James Version* 1975)

" You like potato and I like potahto
You like tomato and I like tomahto
Potato, potahto, tomato, tomahto
Let's call the whole thing off"

— Fred Astaire, in the movie *Shall We Dance*, illustrating also a situation of conflict revolving around pronunciation (Gershwin and Gershwin, 1937)

The concept of *social meaning* was briefly introduced Chapter 1 and is central to this work. Social meaning is a sociolinguistics concept, and as such has emerged in a tradition that is very different from the linguistics and philosophy of language that we've tackled in Chapter 2.

We will go through the origins of this concept in variationist studies and slowly evolve into present-day approaches. The parallels with signaling games, mentioned in 2.3, will be made evident, and we will end on existing formalizations of the concept in game-theoretic pragmatics with SMG.

3.1 WHAT IS SOCIAL MEANING?

A lot of the literature in sociolinguistics has focused on the concept of *variation* and its links with society and the various social groups therein. Among English speakers, for example, you might find groups of people that like to refer to *Solanum lycopersicum* as /təˈmeɪɾoʊ/ while others prefer to refer to it as /tʰəˈmɑːtʰəʊ/. While speakers will recognize those two variants[1] to be the same *word*, "tomato", the pronunciation of each might be seen as a hint regarding the group of people that the person uttering it belongs to.

1 A *linguistic variable* is a set a related forms that, when exchanged, have no influence on the semantic content of the signal (in theory at least) but whose use correlates with a social group. A *variant* is one such form. You can think of them as phonology's *allophones*, where in addition to the phonological features are added social considerations regarding the distribution of the variants within a group. They are defined in Labov (1966b, 1972).

The information about the speaker that can be found in the variation of their speech is what we call "social meaning".

3.1.1 *Social meaning and variation*

To have an idea of what we actually mean by "social meaning", we have to take a step back and see how that notion has been shaped, over decades, in works in sociolinguistics and variation. It has for a long time been treated as secondary, with research mostly focusing on the variation itself and not its interpretation by listeners. Nowadays, however, it has taken a central place in the study of variation, and has been deeply redefined. This more modern approach to social meaning is what we will be interested in. We will mostly be using Eckert (2012).

The idea that the way a language is spoken might give information about the speaker is not new (see the biblical quote above). *Accents* are probably the most obvious example of this, but already when talking about those, it is important not to confuse the works of sociolinguistics with those from lexicographic traditions in the study of variation. For example, Gilliéron and Edmont (1920)[2] is very representative of a geographic approach to variation, listing a number of lexical items and pronunciation variations over the French territory of the time[3]. This geographic approach seeks to map out the many variations of speech over a given territory and attach *variation* to a *place*, linking accent and region of origin.

While accents and dialectal variation are indeed attached to specific places, sociolinguistics is rather interested in how variation is displayed in different social groups. The question is not so much "Do people from Paris and people from Marseille pronounce this word differently?" but rather "Do people from the working class and people from the middle class in Paris pronounce this word differently?". The geographical structuring of language variation is left aside to focus on its value as a social marker.

So when the pronunciation of the diphtong /aɪ/ is discussed in Labov (1963), it is not because we're interested in the existence of this variant on the island of Martha's Vineyard, it is because we're interested in its distribution among Vineyarders. From an ethnographic point of view, what we refer to as *social meaning* could simply be information, in the linguistic signal, about the speaker's geographical origin. When we use it in a sociolinguistics context, however, it is to be understood as information, in the linguistic signal, about the speaker's social group. These can be congruent and are not mutually exclusive (some social groups are found in greater numbers in some territories), but it is the latter we are interested in.

The approach to variation that says that variation *indexes* social groups is called in Eckert (2012) the "First wave" of variation studies. In a few words, the first wave states that in a fixed social world, the many variants in pronunciation or lexicon are simply pointers to specific parts of that social world. We can think of this in truth-conditional semantics terms and see

2 The contents of which can now be found here: http://lig-tdcge.imag.fr/cartodialect5/#/.
3 Avoiding the languages of France that were not of Romance descent, like Breton or Flemish.

the parts of the social world that variants point to as their *extension*. If we envision it that way, the extension of a sociolinguistic variant in the context of First wave variationism is the idealized set of speakers who use that variant. The only thing that variation indexes in that framework is the social group of whoever uses the variant.

Crucially, this approach opposes the *standard* to the *nonstandard* or *vernacular*, and revolves around the idea that any variation at the scale of the idiolect between those two is in fact due to the class system. As explained in Eckert (2012), in the First wave approach, more upper-class individuals use more standard forms, and vernacular forms are more common in the working classes. Situations where upper class people or values are more present will lead to a suppression of the vernacular forms. Among the works that are closer to a First wave approach, we can cite notably Labov (1966a), said to be the starting point of this approach in Eckert (2012), but also the analysis in Rickford (1985) of the use of Creole in Guyana, where a notion of solidarity in nonstandard dialects is underlined – though nuanced by the social class of respondents – and where it is hinted at by interviewees that if one wants to have a job in a managerial position, one should use standard English as much as possible[4].

We can observe one thing here, and it is that the rationale behind the First wave analysis is not very far from what we've said from the ethnographic approaches, except this time, instead of considering a geographical space, we are considering a social space. Crucially, that social space is considered to be *fixed*, and in the cases described in Labov (1966b) is largely reduced to the position of individuals on a working-class/upper-class spectrum. The variants themselves are largely categorized as belonging to *standard* or *vernacular* forms.

The Second wave approach mainly seeks to address this last point by adopting a much more local approach. Instead of focusing simply on working/upper classes, these works focus rather on social groups that have local relevance in the place of study. Interestingly, Labov (1963) is closer to this approach than it is to the First wave, with its opposition between fishing industry and tourism industry. Another key work in this approach is Eckert (1988), with the now famous opposition between *jocks* and *burnouts*.

In a few words: in Eckert (1988), we are shown data about the use of linguistic variants (notably *negative concord*[5]) among groups of high schoolers in the Detroit area. Among these people, some identify with the *jocks* group – referring to people who partake a lot in extra-curricular activities and are generally involved in the local culture, seeking or having some sort of local fame in the high school context and seeing this involvement as generally positive (think football players or cheerleaders) – and others identify more with the *burnouts* – people who value an ideal of street-smartness and urban way of life, inspired by the working class

4 Rickford (1985) and many other works on the subject use an experimental approach called *Matched Guise Technique*, first presented in Lambert (1967). In short, the idea is to present to subjects stimuli that are artificially manipulated so that we end up with pairs of stimuli only differing from one another based on the presence or absence of the variant that we are interested in. Subjects are then asked a number of questions about what they heard and it is assumed that the presence or absence of the variant will account for observed differences. Because of its focus, the technique is adapted to full dialectal variation in Rickford (1985).

5 Sometimes called *double negative*, negative concord in English is a key feature of the African-American vernacular. Think "We don't need no education."

and minorities of the city. These two groups, anything in between and any group beyond (like the *burned out burnout girls*) form the basis of the analysis presented in Eckert (1988), and it becomes apparent that in addition to general clothing style and specific group activities, members of each group adopt specific linguistic variants, and negative concord appears to be used a lot more by youth who identify as *burnouts* or lean towards the *burnout* category than by *jocks*. This leads to the conclusion that linguistic style and variant choice can also be seen as just another brick in the construction of style in general.

Both the First and Second wave approaches offer answers to questions such as "Why do people talk differently according to the context in which they find themselves?". The answers of the First wave, however, do not allow us to make distinctions beyond a *formal/informal* categorization, with formal situations triggering the use of standard forms and informal situations triggering the use of forms that are more natural to the speaker[6]. What the Second wave brings is a much more fine-grained analysis of the variants themselves and how they come into play in given contexts. The locality of the analysis will be the starting point of the Third wave approach, but the Second wave, in and of itself, is not philosophically very different from the First wave with respect to the staticity of the social space. Again, variants are analyzed as indexing specific groups, but this time the variants refer to groups that are much more specified and attached to a context. Saying /tʰəˈmɑːtʰəʊ/ in New York City and saying it in Chicago will not necessarily trigger the same inferences with regards to which group I belong to, and this beyond the notion of *formality*.

Another point put forward by the Second wave approach is that the *vernacular* is no longer seen as the *default*. People actively seek vernacular forms because these forms seem to index positive values (Eckert, 2012; Labov, 1963).

This points towards a different understanding of variant meaning. With First and Second wave approaches, social meaning *per se* was merely thought of as an indexing of *class* (First wave) or *communities* (Second wave). But the work presented in Eckert (1988) already points towards a view of social meaning that is closer to what is now commonly called the Third wave approach, presented notably in Eckert (2008).

3.1.2 *Social meaning in the Third wave*

It is not correct to see the social meaning of variants as being merely their "extension" in terms of the communities to which they refer; such an approach seems to wrongly assume that whenever I use a variant put forward by a specific group, I wish to convey the idea that I am part of that group. This is not the case. In the words of Eckert (2008):

> *"[C]learly, women (and men) are not saying 'I'm a woman' when they use a 'female-led' change, nor are they saying 'I'm not a woman' when they do not. [...] Quantitative generalizations of the sort*

6 These forms can still be standard in case e. g.the speaker is from middle or upper classes.

made in survey studies are important, but exploring the meaning of variation requires that we examine what lies beneath those generalizations. The very fact that the same variables may stratify regularly with multiple categories – e.g. gender, ethnicity, and class – indicates that their meanings are not directly related to these categories but to something that is related to all of them. "

— Eckert, 2008

This operates a shift in what is understood to be "social meaning". The informal definition of it is now expanded to encompass not just the social group of the speaker, but also specific properties or traits that they wish to display. With the First and Second wave approaches, social meaning was derived from variation, but was more or less seen as secondary in the order of events. The meaning of the variants was more or less analyzed as a byproduct of the structure of the social space.

The Third wave approach, on the other hand, does not equate social meaning with social group and does not equate social group with identity. The fragmentation of social meaning is to be linked with usage; speakers actively seek out some variants over others in given situations and according to many parameters. Language variation becomes just another brick in *style* and speakers construct their *identities* through them. The notion of meaning becomes crucial in the analysis, and we can even argue, as is done in Eckert (2012), that the meaning can precede the variant. In other words:

"SOCIAL MEANING BEGETS VARIATION, AND VICE VERSA"

— Acton, 2017

Contrary to the other two waves, the Third focuses a lot more on the meaning of variants, and this meaning is therefore defined further in those works. In Eckert (2008, 2012), the main reference conjured up when trying to define social meaning is Silverstein (2003) and the concept of *indexical order*.

According to this view, the process whereby traits are indexicalized by variants and thereafter used in the construction of styles in conversation goes as follows:

1. First, a population might become salient for any reason, and a distinguishing feature of that population's language may attract attention.
2. That feature is extracted from its linguistic surroundings and indexes membership to that population.
3. It is then used to evoke characteristics or stances associated with that population.
4. Repeated uses of that feature will conventionalize the new sign and make it available for further indexical acts.

Several things can be said about this. First of all, the idea of *indexicalization* is not unlike the concept of *conventional implicature* that we mentioned in Chapter 2, in the sense that beyond the

content of occurrences, some forms might trigger further inferences. The main difference here is that conventional implicatures are used to describe inferences made by listeners about the content of sentences, whereas the notion of indexicality focuses on extra-content information. Still, this is a first hint that the objects of study of pragmatics and of sociolinguistics can be linked.

Another important point here is that the process is *continuous*, meaning that the conventionalization of indexical acts occurs all the time, some appear, some disappear... they come and go following the issues and preferences of a community over time.

A good illustration of this can be found in Woolard (1989, 1991, 2009). Focusing on the social meaning of Catalan in the Barcelona region, where most people are raised bilingual in Catalan and Castilian, the works of Woolard have shown that there are a number of traits that are readily associated to the use of Catalan over Castilian, but it has also shown that, over time, the "solidarity" or *trustworthiness* associated by each speaker to users of their own native language has disappeared, reflecting the normalization of the use of both languages.

The static social meaning of the variants in the first two waves switches to dynamicity, and *the social meaning of a variant* is a notion that can only make sense if a specific context is provided. The meanings indexicalized by a variant at any point in time are called *indexical fields* in Eckert (2012):

> *"[A] constellation of ideologically linked meanings, any region of which can be invoked* in context.*"*[7]
>
> — Eckert, 2012

Several more recent works adopt this notion of social meaning and focus on the indexical fields of variants (Acton, 2017, 2019; Beltrama, 2018; Beltrama and Staum Casasanto, 2017; Burnett, 2017).

Because the Third wave approach focuses on meaning, it has also been taken up by researchers who initially studied meaning and was applied to units beyond the phoneme or morpheme that carry meaning in and of themselves (lexical units, e. g.in Acton, 2019, or just general turns of phrases in Beltrama, 2018). The study of the social meaning of variation is now a lot more present in the study of language meaning more generally.

Another key addition that the Third wave has made to issues of social meaning was to see them as largely *underspecified*. This stems from the fact that any given variant can index many things: which of those things does the speaker want to index in their present utterance? The precise indexical act that is being performed at any point in a conversation is only identifiable through the context of utterance, and social meaning is then seen through the lens of *persona* or *style*.

The idea is as follows: taking into account the context of utterance and their goal in that context, speakers will use the variants that seem more appropriate to construct their identity

7 Emphasis mine.

in that moment accordingly. That identity might be based off of a stereotypical *persona*, characterized as a set of relevant properties. The *jocks* and *burnouts* of Eckert (1988) can be seen, under this light, as personae, being associated with a number of characteristics. The fishermen of Labov (1963) could also be construed as a persona of sorts.

Each individual uses whatever linguistic variant indexes the trait they wish to display and the personae they wish to use in order to construct their own identity in a conversation.

In other words: speakers have at their disposal a number of messages; each message *signals* one or more different traits that speakers wish to be associated with; the messages will be chosen based on the expected interpretations of the audience. If this all sounds familiar, it's because we've characterized communication in very similar terms in section 2.3. With the introduction of the Third wave approach to social meaning, speakers are seen as *agents* with *goals*, and the parallels between some objects of inquiries of pragmatics and of sociolinguistics become evident.

> *"[O]ne finds that meanings are deeply context-sensitive, bound up with ideology, and diverse in kind and source (see e.g.Silverstein, 1976), yet all the while united by general principles of language use and interpretation."*
>
> — *Acton, 2017*

3.2 FORMALIZING SOCIAL MEANING USING SMG

The goal of the SMG framework is exactly this: using tools that have been useful to characterize some issues in pragmatics and cognitive science to try and characterize issues in sociolinguistics. Why would we want to do this? This takes us back to the few remarks that we made about formalization in Chapter 1. The idea is to see the extent to which two seemingly different phenomena can be characterized in the same way, hinting at a common underlying mechanism which, in this case, has been shown to describe accurately processes such as reference determination, scalar implicature inference, and yet other things (see Scontras, Tessler, and Franke, 2018 for examples). Also, in the terms of Burnett (2017):

> *"Formalization can be a very powerful tool for helping us carefully distinguish between different aspects of theoretical proposals and for precisely identifying empirical predictions made by competing analyses."*
>
> — *Burnett, 2017*

But there is more to this. As discussed in Récanati (2010), linguistics has historically evolved towards leaving an ever more marginal place to the study of context and everything that is outside the linguistic signal itself, notably in the study of meaning. This may be due to the structuralist principles discussed in Chapter 2 whereby language (and language meaning) is to be envisioned as a "structure" (Hjelmslev, 1957) for scientific inquiry to be fruitful.

What is argued in Burnett (2017, 2019) is that, following work on social meaning, this prevents us from studying a rich and important part of linguistic meaning. There is a dire need to use tools that can give a larger place to context and the world outside of the signal if we are to properly characterize linguistic meaning. SMG are presented in Burnett (2017, 2019) and rely a lot on the RSA framework presented in section 2.3.2. They also rely heavily on the Third wave conception of social meaning and identity construction to describe the strategic use of linguistic variation in the construction of identity.

The notion of "strategic use of linguistic variation" might seem odd or maybe uncalled for. While it transpires from the way language users are theorized in the Third wave (agents with specific goals), it does not necessarily seems to be appropriate with all kinds of variation. In Burnett (2017), the focus is on the ING variable. In short, this variable concerns the <-ing> word ending in American English and has two main variants: [in] and [iŋ], each associated in the literature with a number of stances and traits (Campbell-Kibler, 2007, 2008). It has been widely studied and is fairly recognizable to the ear, and a strategic use of this (at least to some extent) does not sound far-fetched. In the same work, the *t-release* variable is also mentioned. This one concerns the pronunciation of /t/ and also has two main variants: [ɾ] and [tʰ]. Unlike ING, however, it is much more subtle and it might seem less likely that language users have a conscious control over it. Using the term "strategic" might then sound odd.

The point is, however, that *t-release* has in fact been associated with stances and traits in matched guise experiments[8](Podesva et al., 2015). Whether the control over the variant is conscious is irrelevant, it is very possible that speakers do not have the intention to control how they pronounce this particular variable, maybe it is just a corollary change in their pronunciation that follows from a much broader rule like "Be more formal". In our context, it amounts to the same thing, a preference for a given set of traits/for a given persona will lead to a difference in the use of a variant over another. The strategic choice is done at the level of the persona that one wants to display. The phonetic changes that this brings to a speaker's utterances are the result of that initial choice. This is in keeping with what is meant in Eckert (2012) when it is said that:

> *"It has become clear that patterns of variation do not simply unfold from the speaker's structural position in a system of production, but are part of* the active — stylistic — production of social differentiation." [9]
>
> — *Eckert, 2012*

SMG broadly function like standard RSA models, the key differences that can be found between the two are:

- The social meaning of messages is characterized slightly differently from the denotational meaning of messages that is found in standard RSA.

8 See footnote 4.
9 Emphasis mine.

VARIANT	ECKERT FIELD	ECKERT-MONTAGUE FIELD
/iŋ/	{c, a}	{c, a}, {c, f}, {i, a}
/in/	{i, f}	{i, f}, {c, f}, {i, a}

Table 7: Eckert and Eckert-Montague fields for the ING variant as presented in Burnett (2019). Letters *i, c, a, f* stand for *incompetent, competent, aloof* and *friendly*, respectively.

- The utlity function associated with the expression of personae is philosophically different as well, since the notion of *preference* is much stronger here[10].

Following Third wave variationism, the social meaning of linguistic items consists of the *indexical fields* of those items. This is similar to the way we described the semantic meaning of utterances in terms of possible worlds in 2.3. But again as per Third wave variationism, language users do not attempt to *convey* indexical fields, but styles, or personae, that is sets of stances and properties.

Personae in Burnett (2019) are defined as maximally consistent sets of properties. "Consistent" means that there is no persona such that both a trait t and a trait $\neg$t are in its set of traits. For example, it cannot be the case that a speaker wishes to display a persona that is both *competent* and *incompetent*. "Maximal" means that there is no persona that is simply a subset of another persona, so for example, it cannot be the case that you would have a persona with the properties of being *competent* and *nice* and a persona with the properties of just being *competent*. That second persona must not be taken into account because it is underspecified with regards to the property of *niceness*, which is relevant here[11].

The indexical field of a variant is called in Burnett (2019) its *Eckert field*. These sets of properties indicate to us the personae that can potentially be constructed by those variants. This set of potential personae is called an *Eckert-Montague field*. Table 7 presents the relation between the two concepts as it is presented in Burnett (2019), focusing on the ING variable and the set of properties {competent, incompetent, friendly, aloof} (assuming "friendly" and "aloof" are inconsistent).

Once that has been defined, we can define a standard SMG, where the interpretation of signals will take the guise of pragmatic enrichment and will be very similar to what we have seen for scalar implicatures in section 2.3.2. Let us look at the formal definition given in Burnett (2019).

Definition 3.2.1. Social Meaning Game

A Social Meaning Game is a tuple

$$\langle \{S, L\}, \langle \mathbb{P}, > \rangle, M, C, [\cdot], Pr \rangle.$$

In this tuple:

10 This applies to SMG with persona selection, presented below.

11 Determining which properties – and therefore which personae – are relevant is another question, one that more recent endeavors are attempting to explore. See footnote 13.

- S and L are the Speaker and the Listener.
- $\langle \mathbb{P}, > \rangle$ is the *universe*, where:
  - $\mathbb{P} = \{p_1, \cdots, p_n\}$ is a finite set of properties.
  - $>$ is an irreflexive and asymmetric relation on those properties symbolizing the incompatibility between two properties[12].
- M is a finite set of messages.
- $C : M \to \mathbb{R}$ is a cost function.
- $[\cdot]$ is the indexation relation (see below).
- $Pr \in \Delta(\mathcal{P}(\mathbb{P}))$ is a probability distribution over sets of properties characterizing the prior beliefs of L about S.

Following the definition of $\langle \mathbb{P}, > \rangle$, we can define the set of personae PERS and the personae π that can be found therein, giving a formal definition to what we have mentioned above:

Definition 3.2.2. Persona

π is a persona $\pi \in$ PERS iff:

- $\pi \subseteq \mathbb{P}$ and there are no $p_a, p_b \in \pi$ such that $p_a > p_b$.
  Consistency
- There is no $\pi' \in$ PERS such that $\pi \subset \pi'$.
  Maximality

The *indexation relation* $[\cdot]$ is the counterpart to the interpretation function $[\![\cdot]\!]$. It can be thought of as a function mapping a message to its Eckert field. In the context of SMG however, the Eckert-Montague field representation of indexical fields is preferred. Hence:

Definition 3.2.3. Indexation relation

An indexation is a relation $[\cdot] : M \to \mathcal{P}(\mathbb{P})$, relating messages to subsets of $\mathbb{P}$.

Apart from these slight differences, SMG work like standard RSA and notably use the same solution concept. Burnett (2019) illustrates it with an example focusing on Barack Obama's distribution of ING variants according to different contexts (inspired by e. g.Podesva et al., 2015) and Burnett (2017) does so with the distribution of the same variants according to different contexts for two non-binary individuals, Flynn and Casey (Gratton, 2016).

Choosing the right α parameter (which is necessary to conduct all the steps of pragmatic enrichment present in the RSA solution concept), the model is able to predict similar distributions of variants as those observed in Gratton (2016) for Flynn and Casey.

12 A relation $>$ is said to be *irreflexive* if it is such that for any element x subject to that relation, it cannot be the case that $x > x$.
A relation $>$ is said to be *asymmetric* if it is such that it cannot be the case that both $a > b$ and $b > a$ are true at the same time.
Note that in fact, only asymmetry is needed, because it implies irreflexivity.

PERSONA	π	$\mu(\pi)$
Cool guy	{comp., friendly}	2
Stern leader	{comp., aloof}	1
Doofus	{incomp., friendly}	1
Arrogant asshole	{incomp., aloof}	0

Table 8: A possible μ function presented in Burnett (2019). In the article, it is associated with the informal context of a barbecue.

This picture is however limited. We have mentioned several times already that the Third wave distinguishes itself from the rest of variation studies by putting forward the idea that language users are actively choosing to display some personae over others. While the RSA approach is sufficient to characterize a situation of standard information communication in cooperative contexts, where the right interpretation of a message is not decided by the speaker or the listener, it has to be enriched if we are to describe situations where speakers have preferences that are not dictated by the state of the world.

This leads us to consider SMG with persona selection, which are identical to SMG except for the addition of a preference function $\mu : \text{PERS} \to \mathbb{R}$, mapping each possible persona to the value that S gives it in the context of utterance. The game is otherwise identical.

Definition 3.2.4. Social Meaning Game with persona selection

A Social Meaning Game with persona selection is a tuple

$$\langle \{S, L\}, \langle \mathbb{P}, > \rangle, M, C, [\cdot], Pr, \mu \rangle$$

The μ parameter is better understood as being derived from a broader game in which the linguistic interaction is just a part. It is argued in Burnett (2019) that it makes sense to use it in some interactions, focusing on the example of Obama's use of ING variants. This view makes sense in general in the context of politics, where the interests of speakers go beyond the accurate communication of content and information about themselves. We will come back to this, since it also makes a lot of sense to assume similar preference relations when discussing dogwhistles, where acquiring listeners' approval is a goal that supercedes the accurate communication of information.

The addition of μ allows us to make sense of idiolectal variation, the difference in speech patterns observed at the scale of a single individual according to the context of utterance. With proper μ functions and α tuning, Burnett (2019) manages to propose a model that correctly predicts that Barack Obama's use of the two variants /in/ and /iŋ/ will vary according to the context. Tables 8 and 9 present some possible μ functions for illustration.

The solution concept for regular SMG is almost the same as what is found in standard RSA (2.3.2), with the t parameters for states of the world being substituted with π parameters for personae and the interpretation function $[\![\cdot]\!]$ being substituted with the indexation relation $[\cdot]$.

PERSONA	π	$\mu(\pi)$
Cool guy	{comp., friendly}	1
Stern leader	{comp., aloof}	2
Doofus	{incomp., friendly}	0
Arrogant asshole	{incomp., aloof}	0

Table 9: A different possible μ function. We can for example envision this to be more associated with a press conference on serious matters.

Taking the listener prior beliefs presented in Table 10, a standard SMG approach to this with $\alpha = 6$ leads to the following results:

$$P_S([\text{in}]|\text{Cool guy}) \approx 0.69$$
$$P_S([\text{iŋ}]|\text{Cool guy}) \approx 0.31$$

The frequentist interpretation of this given in Burnett (2019) is that in case Obama wants to convey the Cool guy persona to his audience, then he will use the [in] variant 69% of the time. The effects on the listener in the model, using the priors in Table 10, are found in Table 11. The paper gives all the details of the computations and the interpretation of those results, but the part that interests us most here is to see the effect of μ on this.

To do this, we have to introduce another choice rule, also based on soft-max, that takes into account μ:

$$P_{\text{PERS}}(\pi; \mu) = \frac{exp(\alpha' * \mu(\pi))}{\sum_{\pi \in \text{PERS}} exp(\alpha' * \mu(\pi'))}$$

Including this allows us to have an estimation of how much Obama will use each variant:

$$\mathfrak{P}_S(m) = \sum_{\pi} P_{\text{PERS}}(\pi; \mu) * P_S(m|\pi)$$

Setting $\alpha, \alpha' = 6$ and using the values for μ found in Table 8, we obtain approximately the same results that we had before for the distribution of variants (since the Cool guy persona is preferred), but we have now implemented directly the idea of context-based idiolectal variation using preferences. So for example, if we use the μ presented in Table 9, we will instead have the following results:

$$\mathfrak{P}_{\text{Obama}}([\text{in}]) \approx 0.001$$
$$\mathfrak{P}_{\text{Obama}}([\text{iŋ}]) \approx 0.999$$

PERSONA	Stern leader	Cool guy	Asshole	Doofus
π	{c, a}	{c, f}	{i, a}	{i, f}
$Pr(\pi)$	0.30	0.20	0.30	0.20

Table 10: Possible listener priors presented in Burnett (2019).

PERSONA	Stern leader	Cool guy	Asshole	Doofus
/in/	0.375	0.25	0.375	0
/iŋ/	0	0.286	0.428	0.286

Table 11: Listener interpretations presented in Burnett (2019) after hearing either of the two variants, based on the priors in Table 10.

These results can be made more or less dramatic by tuning the α, α' parameters and choosing different μ, but the point is that the preferences of the speaker hugely impact the results in a SMG with persona selection.

Burnett (in press) proposes an overall review of this approach and integrates it into the *conceptual spaces* framework (Gärdenfors, 2004), offering a richer representation of the space of personae. The approach that is put forward in Heather Burnett's work is an important turn in the study of these phenomena, because it treats them in a way that is almost identical to the way that phenomena in content-focused pragmatics are treated, bridging the gap between the vastly informal and data-centered sociolinguistic studies and the heavily formal, theory-focused issues of semantics and pragmatics, unifying seemingly vastly different processes under a similar general cognitive mechanism.

We can however notice that this approach is not entirely in keeping with the Third wave agenda in the sense that this modeling of the situation is at its core cognitive (and relies on modeling techniques found in cognitive science), while one of the intentions behind the Third wave approach was to go beyond individual-based cognitive approaches of individual variation and towards a properly social approach. What we have here is a model that can be applied to both speakers and listeners, but that also relies on internal representations that are chosen fairly arbitrarily. The personae themselves are set manually, so is the μ function, and a formal representation of the social space and the emergence and disappearance of personae is lacking.

The brand new framework of *Pragmatic Sociolinguistics*[13] seeks to answer these shortcomings by proposing a model of the social world based on *pragmatic sociology* (Boltanski and Thévenot, 1987), according to which contexts can be defined that will lead to different valuations of properties (hence personae), putting these individual cognitive models back into the social world.

13 Carried by the SMIC project (ERC grant N°850539). More information about the project, its members and their work can be found here: http://www.socialmeaning.eu/.

3.3 CONCLUDING REMARKS

This section tackled the elusive notion of *social meaning* and presented another way in which individuals can communicate information beyond the content of a linguistic signal, while still relying on that signal. More precisely, this argues that even though it is largely context-determined, social meaning is a part of linguistic meaning and can in fact be derived much like implicatures can be derived (in that case, using very similar tools).

As underlined, the framework put forward by SMG is just a first step in discussing social meaning and the way it is transmitted from speaker to listener. Nonetheless, this shows that we can use RSA-inspired tooling to derive other sorts of meaning than simple content-focused meaning. It underlines the similarities between the communication and construction of *identity* and the communication of *content*.

How about when we try to articulate both? Is there content that can be determined by identity? We argue that there is, and this is the first step towards the analysis of dogwhistles that we will propose in the next chapter. Taking inspiration from RSA and SMG, we will propose a model that relies on social meaning information to derive content information and argue that this is one of the core mechanisms of dogwhistle politics.

Part III

DOGWHISTLES AND IDENTITY-BASED INTERPRETATION

This part focuses specifically on the case of *dogwhistles* as a case of identity-based interpretative variation, first presenting a definition of dogwhistles before linking this concept to the ideas viewed in the previous parts and proposing a formalization of the phenomenon.

4

DEFINING DOGWHISTLES

"SIR TOBY BELCH:

Is't possible?

FABIAN:

If this were played upon a stage now, I could condemn it as an improbable fiction."

— Fabian making a comment that only the audience can truly understand (*Twelfth Night*, 3.4) (Shakespeare, 1623/1910)

"The ACME 'Silent' Dog Whistle is precision engineered to produce a range of fundamental frequencies between 5,400 Hz and 12,800 Hz. Although at this level, the whistle is insignificant to human hearing, a dog, whose ears are far more sensitive to high frequencies, will instantly become more alert to the sound of this whistle than to any other, even one of louder and lower pitch."

— User manual of an actual dog whistle (J. Hudson & Co. Ltd., 1991)

A dog whistle is a special kind of whistle that only emits sounds in frequencies that can be heard by dogs, but not by humans. *Dogwhistle politics* is a term generally used to describe situations where a politician sends seemingly innocuous messages to their audience while also sending less acceptable messages to a subset of their audience – the "dogs" in the analogy.

In more precise terms:

"'Dog whistle politics' is a way of sending a message to certain potential supporters in such a way as to make it inaudible to others whom it might alienate or deniable for still others who would find any explicit appeal along those lines offensive."

— (Goodin and Saward, 2005)

This definition will be taken as a starting point in the analysis of the phenomenon presented here. It gives a good general idea of what is meant by the term "dogwhistle", although we will see that it has to be – and has been – enriched to give a proper account of these discourse moves. This chapter will focus on the more common definitions of dogwhistles, the history of the concept and its use, but also what can be made of it in linguistics and philosophy of language.

Specifically, a lot of pragmatics after Grice (1967) has revolved around the idea of *cooperation* and the truthful exchange of information. This is not what dogwhistles are about. In fact it

is not what political language is about, most of the time. This is one reason to try and study dogwhistles from a linguistics perspective: broadening the subject-matter of pragmatics, one object of study at a time, and attempt to evade "ideal language theory" (Beaver and Stanley, 2018).

4.1 DOGWHISTLES AND POLITICAL STRATEGY

In Goodin and Saward (2005) it is argued that dogwhistling is merely the revival of a more ancient practice: "whistle-stop campaigning". This can basically be understood as a candidate for a given office telling different things to different communities according to their local interests. Because of segmented news markets, candidates would do this in complete confidence that no one would ever notice the discrepancies. With the advent of wider means of communication, politicians have had to rely on other strategies to keep adressing diverse audiences in diverse terms, but the end goal is the same: elections are won with votes; any strategy that leads to a maximization of the amount of votes is useful in winning an election. Gathering the support of people from many different backgrounds is done more easily if each group of people hears what they want to hear.

Dogwhistling, in principle, allows politicians to discuss various topics with various members of one given audience, using a unique message. The topic has been studied mostly by political science and philosophy, with many arguing that it presents a threat to democracy, either in favoring mandates that are weaker (on account of being more opaque regarding their goals[1]) or in jeopardizing the very foundations of reasonableness and reasonable discussion, taken as pillars of democratic societies (Stanley, 2015). One thing seems certain: the practice has acquired a name and bad press; audiences notice it.

The fact that it is noticed does not however mean that it is intentional. In fact, the expression "Dog Whistle effect", usually attributed to Richard Morin (in Safire, 2008), refers to the *unintended* interpretations of polling questions. The analyses presented in Saul (2018a,b, 2019) also insist on the *uninentional* aspect of dogwhistles. The term "dog-whistle" itself however, does not have any particularly unintentional color to it either; the whistle is blown by an intentional agent for specific goals. As is the case with actual dog whistles, it has become obvious that dogwhistle speech is *to some extent* intentional. Several actors of the political landscape have admitted to using strategies close to dogwhistling, especially among right-wing parties[2]:

1 "[The] difficulty of inferring a mandate from mixed-message politics is redoubled in cases of dog whistle politics. Imagine a particularly extreme example. A conservative party dog-whistles an encouraging message to racists that its own traditional supporters would instantly repudiate. It wins the ensuing election. Half its voters voted for it purely because of its (coded) support for racist policies; half voted for it purely because of its traditionally decent policies on race. Clearly, the party won a majority; clearly, it has a mandate to rule. But under those circumstances, it equally clearly could not claim a policy mandate to pursue either of the two contradictory policies that won it its votes." (Goodin and Saward, 2005)

2 Importantly, however, dogwhistling is not specifically a right-wing tactic. See Haney-Lopez (2014, p.4): "Dog whistling has no particular political valence, occurring on the right and the left, nor is it especially uncommon or troubling in and of itself. Given a diverse public segmented by widely different priorities, it is entirely predictable that politicians would look for shrouded ways to address divergent audiences."

> *"There's a difference between selling out your ideas and selling your ideas, and the British National Party isn't about selling out its ideas, which are your ideas too, but we are determined now to sell them, and that means basically to use the* saleable words*, as I say, freedom, security, identity, democracy. Nobody can criticise them. Nobody can come at you and attack you on those ideas. They are* saleable*.[...]"*
>
> *"Once we're in a position where we control the British broadcasting media, then perhaps one day the British people might change their mind and say, 'yes, every last one must go'. But if you hold that out as your sole aim to start with, you're not going to get anywhere. So,* instead of talking about racial purity, we talk about identity.*"*[3]
>
> — Nick Griffin, head of the British National Party (BNP), adressing the American Friends of the BNP in 2000 (Corbin, 2001)

While this does not directly refer to dogwhistle politics, it seems to be of similar inspiration and underlines one of the potential long-term strategies/effects associated with dogwhistles, which we will call "opening the window" – this is explored a bit more in Chapter 9. In any case, Griffin underlines here the interest, for some political groups, to have a duplicitous discourse: gather as much support as possible. Griffin's idea here is basically that if it wants to implement its more radical policies, the BNP first has to sell watered-down versions of these policies in order to educate the public into accepting ideas that, at face-value, sound unacceptable.

Politically, this is close to dogwhistling, in the sense that the ideas that are defended by BNP officials are sugarcoated to suit the public's taste, while remaining transparent in their endgoal to historical supporters within the party. Philosophically, dogwhistles go a bit further in saying that the language used by BNP officials *means* different things to historical supporters and to the general public. Those "*saleable words*", in dogwhistle theory, have several meanings, a *saleable* one and a *hidden* one. When Griffin says: "instead of talking about racial purity, we talk about identity", one reading of this using a dogwhistle approach is that in a BNP context, "racial purity" and "identity" mean the same thing.

Remember from Chapter 2 that the *meaning* of a word is not necessarily simply its extension in terms of the possible worlds that it denotes. Some of the meaning of words and utterances are conventional, and more importantly, we've seen with the signaling game approach to communication that *meaning* can be seen as the result of recursive thinking on the part of a listener and a speaker, notably in cases of ambiguity. It is not implausible at all that any given message is understood differently according to the person who hears it, and all the more so when value-laden, yet vague, words such as *freedom* or *identity* are concerned. The work presented in Lindgren (2018) and Lindgren and Naurin (2017) explores this further.

This can already lead us towards considerations for pragmatics in general: a Gricean approach focusing on the communication of accurate statements about the world should theoret-

3 Emphases mine.

ically lead to those vague words not being used very often, if at all, since they possibly favor a host of unwanted, sometimes opposing, interpretations. If we step out of cooperativeness for a moment though, it seems very obvious why using those words is politically very interesting.

One reason behind doing this is, e. g., the following, in Stanley (2018):

> *"It's hard to advance a policy that will harm a large group of people in straightforward terms. The role of political propaganda is to conceal politicians' or political movements' clearly problematic goals by masking them with ideals that are widely accepted."*
>
> — (Stanley, 2018)

But it is important to remember that the main reason behind this strategy remains: gathering as many votes as possible, and that it is not necessarily used to hide an unspeakable agenda so much as it's being used to persuade voters of their inclination towards one candidate or another.

4.1.1 *Race and dogwhistles in the USA*

While the practice is not limited to either the USA or race-based discourse[4], this is what a lot of the analyses of the phenomenon first focused on, especially after the analysis found in Mendelberg (2001).

Griffin's quote is not the only example of open discussion of similar political strategies. Another much discussed quote is this one, from then member of the Reagan administration Lee Atwater, describing a point since then analyzed as being part and parcel of the so-called *Southern strategy* of the United States' Republican Party:

> *"You start out in 1954 by saying, "N*****, n*****, n*****". By 1968 you can't say "n*****" – that hurts you. Backfires. So you say stuff like forced busing, states' rights and all that stuff. You're getting so abstract now [that] you're talking about cutting taxes, and all these things you're talking about are totally economic things and a byproduct of them is [that]* blacks get hurt worse than whites. And subconsciously maybe that is part of it. *I'm not saying that. But I'm saying that if it is getting that abstract, and that coded, that we are doing away with the racial problem one way or the other. You follow me – because obviously sitting around saying, "We want to cut this", is much more abstract than even the busing thing, and a hell of a lot more abstract than "N*****, n*****". So, any way you look at it, race is coming on the backbone."*[5]
>
> — Lee Atwater, quoted in Lamis (1990)

4 The concept has also been applied, among other things, to the use of Twitter by white supremacists (Bhat and Klein, 2020), anti-elite discourse (Bonikowski and Zhang, 2020), sanctuary cities in the USA (Lasch, 2016) or the interpretation of slogans during the Brexit campaign (Saul, 2018b).

5 Emphases mine.

Quotes such as this one have participated in the representation that is now made of the practice of dogwhistle politics in philosophy (Saul, 2018a; Stanley, 2015) and social psychology (Mendelberg, 2001) in particular.

The analysis found in Mendelberg (2001) is not language-centric; one of her most prominent examples is that of the Willie Horton ad during the 1988 presidential campaign. In short, the ad depicts William R. Horton, a criminal of the state of Massachusetts granted a weekend furlough from which he did not return; he then committed assault, robbery and rape. The mugshot of Horton, a black man, is presented on the screen while a voiceover comments on his actions, blaming them on Michael Dukakis, then candidate to the presidential election, facing George H. W. Bush.

The ad was hugely successful, but already at the time was accused of race-baiting due to the presence of the picture of Horton on-screen (there is no explicit mention of race in the voiceover). One analysis that is made of this (and was then made by Jesse Jackson) is that the presence of the picture allowed unconscious racial resentment from white people towards black people to become operational, leading to a stronger support in favor of the candidate that called upon this resentment. The mobilization of known anti-black stereotypes in the minds of voters would have had an influence on their view of Bush's *tough on crime* policies. One (unverifiable) corollary to this argument is that had the criminal been white, the ad would not have been as effective.

If there is racial resentment in the population however, it would certainly be more efficient for the ad promoters to make explicit mentions of race. It is argued in Mendelberg (2001) that this would not have worked due to what is called the "norm of racial equality". In short, being viewed as racist is generally frowned upon in post Civil Rights USA, and the underlying, openly racist argument made in the ad regarding black people and criminal behavior would be irreceivable. The idea behind the norm is that in most cases, people would consciously refuse arguments that would classify them as racist if they approved of them, but because there is still some degree of underlying racial resentment, playing indirectly on an anti-black sentiment can still be effective. For the one who makes it as well as for the one who accepts it, there is a degree of *plausible deniability* here: I am not racist because I am not defending a racist argument; I am defending the idea that Dukakis is soft on crime.

The reason invoked in Mendelberg (2001) to justify that the ad was indeed relying on an unconscious racial resentment and to illustrate the existence of a norm of racial equality is that once Jesse Jackson did comment on the racial undertones of the ad and his commentary was in turn commented upon (mostly in disagreement) in the media, then the voting intentions shifted again, as if the ad was suddenly less successful, thereby showing that some voters, once made aware of the racially problematic content of the ad, decided against its message.

The term "dogwhistle" is not found in Mendelberg (2001), but this analysis is very compatible with the initial definition we gave of the phenomenon. A political party used ambiguous

messaging to signal some degree of racial bias to an audience that was sensitive to it in order to gather more support in such a way that it could deny having done so.

The concept of a "norm of racial equality" has been taken up notably in Stanley (2015) and allows to explain the following:

- The fact that the Bush campaign could not hold openly racist views.
- The fact that once made evident that there was an attempt at circumventing this norm in the ad, its effectiveness was seemingly lower.

Importantly, this example underlines the importance of a "norm", a possibly unwritten rule of discourse preventing the discussion of certain topics in a certain way. If discussing the taboo topic would profit to some political entity, then ways will be found to circumvent the norm.

In the wake of the work presented in Mendelberg (2001), a number of empirical studies have been made to measure how effective this strategy actually was. In Valentino, Hutchings, and White (2002), the effect of racial pictorial representations in political ads is experimentally tested via the manipulation of a George W. Bush campaign ad. The results of their experiments show that the presence of pictorial racial cues seem to prime racial attitudes, in keeping with the approach in Mendelberg (2001). Beyond pictures, it is shown in Hurwitz and Peffley (2005) that the phrase "inner city" does have an effect on the support offered to some policies, and that this effect correlates with participants' racial attitudes, when compared with messages without implicit racial content.

"Inner city" is a common example that has been taken up in e. g.Saul (2018a) and Henderson and McCready (2019b), notably in the following form:

(15) a. We have got this tailspin of culture, in our *inner cities* in particular, of men not working and just generations of men not even thinking about working or learning the value and the culture of work.

b. We have got this tailspin of culture, in our *African-American neighborhoods* in particular, of men not working and just generations of men not even thinking about working or learning the value and the culture of work.

Emphases mine

(15a) is a sentence uttered by Representative Paul Ryan on the radio program "Morning in America", hosted by Bill Bennett, in 2014. (15b) is one interpretation of the meaning of this sentence making a possible dogwhistled content overt. Whether such a *content* is indeed conveyed in the utterance is disputed (Khoo, 2017) and will be discussed below.

In White (2007), it is shown that the effects observed with messages with implicit racial content disappear when the racial content is made explicit[6] with white participants. A different

6 For results going against the idea that implicit racial appeals may be more effective than explicit racial appeals, see Huber and Lapinski (2006, 2008), but also Valentino, Neuner, and Vandenbroek (2018) for an updated view taking into account the changing political landscape.

pattern is observed in black participants, with whom implicit racial messages do not necessarily trigger racial identification, but explicit messages do. In Wetts and Willer (2019), it is shown that the effect of implicit racial appeals can be further explained by political affiliation, with liberals scoring higher on racial resentment scales being much more likely to have their support of certain policies influenced by racial cues (interestingly, both implicit and explicit[7]).

The effect of those coded racial appeals is attested, they seem to be an effective way to indirectly play on the racial resentment felt by some white voters while remaining (mostly) safe from accusations of racism. A thorough review of dogwhistle racism and its effect on the political discourse in the USA can be found in Haney-Lopez (2014), where we find the following, key point:

> *"At root, the 'racism' in dog whistle racism is the 'strategy' in the Southern Strategy"*
>
> — (Haney-Lopez, 2014, p.113)

According to Haney-Lopez (2014), stirring racial resentment is not limited to the Republican party[8], nor does it necessarily stem from particularly racist views. While it seems hard to deny that some of the promoters of that approach to political communication did have views on race that would definitely go against the norm of racial equality[9], this is not necessarily the best way to explain the use of this kind of discourse:

> *"They may have harbored tainted beliefs, but racial animosity did not drive their actions. Instead they concentrated hard, weighing and sifting, to figure out how they could most effectively gain votes."*
>
> — (Haney-Lopez, 2014, p.48)

This political strategy is illustrated in what is called "Punch, Parry, Kick" in Haney-Lopez (2014, p.129):

- *Punch*: Using coded race talk in the form of a dogwhistle.
- *Parry*: Refuse to see the connection between the comments and race.
- *Kick*: Accuse the opposing party of mentioning race in a conversation from which it was absent.

The strategy has been very effective for Republicans in the USA. Fully acknowledging the strategic dimension of dogwhistles forces us to highlight one thing that is not made explicit in the definition in Goodin and Saward (2005). While dogwhistles can be presented as a way

7 This can be explained by a change of attitudes regarding norms of racial discourse, which is further explored in 9.2.2.

8 "Clinton's equation of 'hardworking Americans' with 'white Americans' struck many as a form of intentional dog whistling" (Haney-Lopez, 2014, p.111), although it is acknowledged that the instigators of such methods and the most prominent users were in fact Republicans, in particular Barry Goldwater, Richard Nixon and Ronald Reagan.

9 "President emphasized that you have to face that the whole [welfare] problem is really the blacks. The key is to devise a system that recognizes this, while not appearing to... Pointed out that there has never in history been an adequate black nation, and they are the only race of which this is true. Says Africa is hopeless."H. R. Haldeman, Richard Nixon's chief of staff, quoted in Feagin, Vera, and Batur, 2001

for politicians to secure the support of more radical fringes of their party while appearing "acceptable" to others, it can also be used in other ways. For example, it can be a way for politicians to try and convince the fringe voters who typically do not vote for them to vote for them. This might be what was behind Bill Clinton's use of religious language, for example (see the next section). What the experimental results in Wetts and Willer (2019) show however, is that the use of dogwhistle strategies has more effect on the centermost voters (in that case, self-identified liberals displaying a higher degree of racial resentment).

What seems likely then, and especially applies to dogwhistle racism in the USA, is that dogwhistles are not so much used to keep the fringe voters – in a bipartisan system like that of the USA, it seems unlikely that they would vote for any other party than the one they typically identify with anyway[10] – but rather to shift the voting intention of those in the middle who would be receptive to unconscious biases and/or single-issue voters.

At the end of the day, the goal of the American politicians participating in dogwhistle racism is to gain the support of as many people as possible. That is also the goal put forward by Nick Griffin. But we can see that even though the goal of the BNP in the early 2000 and that of the Republican Party after the Civil Rights movement seem to align at first glance, there is a key difference between the two. The goal of the BNP is to present its ideas in such a way that people will agree with them and gradually change their minds about more extreme positions; the goal of the Republican Party is to use the existing racial resentment in the US population to get elected. In the reading of the situation found in Haney-Lopez (2014), the election *is* the end game for dogwhistling politicians in the US; in the view presented by Griffin, it is but a stepping stone.

4.1.2 *Religious dogwhistles in the USA*

Religious dogwhistling is another area where dogwhistles have been extensively studied. Remember (3a), repeated here:

(16) Yet there's power, *wonder-working power*, in the goodness and idealism and faith of the American people.

As stated in Chapter 1, this is a very commonly cited example of dogwhistling, and specifically religious dogwhistling. While dogwhistle racism focused on stirring racial resentment in the audience, the goal of such examples of religious dogwhistling is claimed to be rather different. In this case specifically, *intentionality* is much more prominent. As stated in Chapter 1, the phrase "wonder-working power" comes from the hymn *There is Power in the Blood*, and this instance of dogwhistling is much closer to the image of a *secret handshake*.

10 Though it could very much also be the case that the dogwhistling is intended to convince fringe voters who do not usually vote to vote.

The idea here is that the utterer of a sentence such as (16) would signal their identity as an Evangelical Christian (this community being, supposedly, the only one to use such turns of phrases) to Evangelical Christians, unbeknownst to the rest of the audience (who is unfamiliar to the reference and supposedly only sees in it standard conservative speech). The strategic intention here is also attested through testimony, see e. g.:

> *"We inserted snippets of old hymns in economic sections – such as "the solid rock of economic principles." We threw in a few obscure turns of phrase known clearly to any evangelical, yet unlikely to be noticed by anyone else, even Kemp. Phrases like "narrow is the path of wealth." It was code."*
>
> — (Kuo, 2006, p. 59)

Evangelical David Kuo is here mentioning an episode where himself and one of his colleagues wrote a speech for politician Jack Kemp that had to be given in front of the Southern Baptist Convention. Kuo later served as special assistant to George W. Bush. In Kuo (2006), he underlines the importance of Evangelicals as a political force in the USA as well as the attempts of politicians to tap into that force[11], in some cases successfully, but in the end to the detriment of the broader Evangelical agenda. He concludes his book on a bitter note, urging Christians to step down from the political scene, for a time.

As was the case with dogwhistle racism, there have been experimental studies testing the effectiveness of subtle religious cues on gathering political support in the context of the USA. In Calfano and Djupe (2009), it is shown that the use of subtle religious cues allow Evangelicals and mainline Protestants to identify the candidates as Republicans and does trigger higher support among Evangelicals, but not among mainline Protestants, for whom the presence or absence or religious cues has no effect on support. The cues have no effect either on party identification or on support for Catholic voters. In addition to these results, Albertson (2015) shows that obvious religious cues in discourse come with a cost in support. While there is no "shameful content" in those appeals, the role of religion in public and political discourse is very disputed and people outside the religious ingroup targeted by the cue[12] may be rejecting obvious religious cues because of this.

As was the case with dogwhistle racism, we have signals that carry a content that has an effect on a subset of the audience while being ignored by the rest[13]. But if all of this sounds different from what we saw with dogwhistle racism, it's because it is, and that difference has been used for classification (Henderson and McCready, 2019b; Saul, 2018a).

The differences lie mainly in the following:

1. The dogwhistled content seems to be a form of *social meaning content*, which did not seem to be the case with dogwhistle racism (at least at first glance).

11 Notably then POTUS Bill Clinton, also mentioned in Kuo (2006).
12 In the case of Albertson (2015), the ingroup in question were Pentecostals.
13 Or at least having no-effect on their voting intentions.

2. The dogwhistled content does not seem to be as deniable, meaning it would be a lot harder for, e. g., George W. Bush to deny that he's using an Evangelical turn of phrase in uttering (16) than it is for representative Paul Ryan to deny that he is cazlling forth unconscious racial resentment in uttering (15a).

3. The dogwhistled content does not seem to be particularly offensive.

Let's tackle the third point first: because a lot of the work on dogwhistles has historically carried on racism, the notion has connotations of shamefulness: you dogwhistle content that would otherwise be unacceptable to the audience; an Evangelical upbringing is not particularly unacceptable in a US context. This point is what makes e. g.Mark Liberman say that the specific example in (16) maybe should not be called dogwhistle[14]. However, the shamefulness of dogwhistled content is not a necessary condition of the dogwhistle as we understand it here. The point here is that some content of the utterance is accessible to some people in the audience and not to others, which is the case with (16). If there is a strategic interest to using that kind of messaging instead of straightforward, clear messaging (as in (3b)), then it will be done.

This leads us to a more profound point: any discourse containing cultural references can then be considered to be akin to dogwhistles as long as a part of the audience does not *get it*. This is a position that is held in, e. g., Saul (2018a) (discussing work by Kimberly Witten):

> *"Although the main interest of dogwhistles lies in their political use, Witten rightly argues that the concept applies more broadly. As a parent, I was shocked to revisit some of my favourite childhood entertainments and see much that I had missed as a child. Watching Bugs Bunny with my small son, I was surprised to see references to old movies that children couldn't be expected to know, and even more surprised to see that one of these was* Last Tango in Paris*. Finding these references of course made the endless re-viewings less tedious. And, of course, this was the intent of their makers."*
>
> — (Saul, 2018b)

This underlines also an earlier quote in Haney-Lopez (2014): dogwhistling is not "especially uncommon or troubling in and of itself".

Regarding the other two points, they have been useful in attempts at dogwhistle classification.

4.1.3 *Categories of dogwhistles*

It appears from what we have seen so far that the objects that are referred to using the term "dogwhistle" are many-faceted, and while they all share the common key point of being interpreted differently according to the listeners unbeknownst to them (or at least to some of

14 See https://languagelog.ldc.upenn.edu/nll/?p=50278

them), they can be differentiated at least according to the content that they communicate and the way that they are used and received.

We will focus on two classifications of dogwhistles, the first one is presented in Saul (2018a) and stems from the tradition of philosophy of language and speech act theory, and the second, presented in Henderson and McCready (2019b), is rooted in the tradition of formal linguistics. Both classifications will be criticized and will be useful in giving a more precise definition of dogwhistles as well as stating *which* dogwhistles in particular interest us.

4.1.3.1 *In Saul (2018a)*

In Saul (2018a), dogwhistles are classified according to two axes: their *intentionality* and their *explicitness*. This second distinction only applies to intentional dogwhistles, leaving us with three categories.

1. *Intentional dogwhistles*
    a) *Explicit intentional dogwhistles*: This category would be the most prototypical instance of dogwhistles. It encompasses all messages designed with the intent to allow several possible interpretations, with some of these being hidden to a part of the audience. Importantly, they are *explicit* in the sense that both the speaker and the listeners of the ingroup are *aware* of the hidden interpretation. These truly qualify as *coded messages* in that sense. (16) qualifies as an explicit intentional dogwhistle. Another example discussed in Saul (2018a) is that of the *Dred Scott* decision. The decision itself stated that no black person in the USA could be a citizen, free or slave. Politicians such as George W. Bush have repeatedly stated in public their opposition to the Dred Scott decision. This opposition is uncontroversial, but according to Saul (2018a) is in fact dogwhistling a much more controversial opposition to another Supreme Court decision: *Roe v. Wade*, which protects the right to abortion.
    b) *Implicit intentional dogwhistles*: This category is more subtle. It contains all occurrences of speech that trigger ingroup thinking and have an effect on political support by framing the debate in a certain light, activating some particular modes of thinking about an issue; importantly, the targeted ingroup is not aware of any particular coded message here. The use of "inner city" to trigger racial ingroup thinking among racially resentful white people in the USA is an example of this (see (15a)). According to what we find in Saul (2018a), the term "inner city" does not have a hidden racial meaning in the propositional sense, but its use plays on racial stereotypes, and if the listeners did have some racial resentment, then this racial resentment will guide their thinking on the issue being discussed, à la Mendelberg (2001).
2. *Unintentional dogwhistles*: Unintentional dogwhistles are, in a way, a consequence of the existence of implicit dogwhistles. If the audience is mostly unaware of the dogwhistley

part of a implicit dogwhistle, they might very well repeat it while being unaware that they are participating in the manipulation that it involves. This category covers notably all instances of repeated discourse, or terms entering regular language use but not losing their dogwhistle properties. This has the profound implication that political manipulation can occur without intent on the part of the speaker.

The discussion in Saul (2018a) underlines that implicit intentional dogwhistles are a lot harder to describe properly and capture, especially in more traditional approaches in language sciences:

> *"Implicit intentional dogwhistles are substantially more challenging to capture. There are two key reasons for this. First, what is dogwhistled is not a particular proposition. Instead, certain pre-existing attitudes are brought to salience. This means that* any theory relying on the communication (via semantics or pragmatics) of a particular proposition will fail. *Second, this occurs outside of consciousness. Crucially, when an audience becomes conscious of the dogwhistle, it fails to achieve its intended effect. Success of an implicit intentional dogwhistle, then – unlike most communicative acts – depends on the audience not recognizing the speaker's intention.* Any theory which includes the idea that uptake (recognition of the speaker's intention) is required for success will fail entirely as a way of accommodating implicit dogwhistles. *Implicit intentional dogwhistles only succeed where uptake is absent; uptake prevents such dogwhistles from being effective."*[15]
>
> — (Saul, 2018a)

Although we will mostly focus on explicit intentional dogwhistles, this position on implicit dogwhistles can be discussed. There are two main things defended here, the first is that the dogwhistled content of covert dogwhistles is not propositional in nature. This means that whatever supplementary meaning the dogwhistle term carries that triggers the stereotypical thinking associated with, e. g., dogwhistle racism is not formulated as a proposition, and any interpretation of the sentence that adds propositional meaning to it goes beyond the sentence itself. I take it to mean that the explicit interpretation of (15a) presented in (15b) is taken to be inadequate, since it is not propositionally equivalent to (15a). However, we have seen in Chapter 2 that we can think of mechanisms whereby propositional content will be inferred by listeners even though it is not explicitly present in the speaker's utterance: this is what happens with scalar implicatures, where "some but not all" is not equivalent to "some"[16], yet is inferred by listeners in the right context. In the appropriate context, listeners will infer from an utterance additional propositional content that, in a RSA framework, they will ascribe to speaker intentions. Importantly, that the speaker intention be verified is not necessary for this to happen; given a set of alternative messages and rationality assumptions about the speaker, this supplementary propositional content will be inferred no matter what the *actual* intentions

15 Emphases mine.

16 But the states of the world covered by "some" include those covered by "some but not all", which would not necessarily apply in the case of a dogwhistle like "inner city": it is not necessarily the case that all "African-American neighborhoods" are in fact found in "inner cities".

of the speaker are. This discussion will be taken up in 4.2, but let's not throw the propositional baby out with the dogwhistle bathwater.

In case the argument put forward is that the dogwhistled content *cannot* possibly be expressed in a propositional form, we have seen in 3.2 that we could use frameworks similar to what is used for the derivation of scalar implicatures to derive the socially meaningful content of utterances, e. g.using SMG. Although social meaning is not generally understood to be propositional, we could treat it in a way similar to how propositional meaning is treated in RSA and end up with a convincing representation of the cognitive mechanisms behind the processing of socially meaningful utterances.

In other words: even if we do not consider dogwhistled content to be propositional (which is very defendable in itself), we can still treat it as propositional from a formal point of view to try and devise models to describe the interpretation of dogwhistles.

The second point that is addressed here is that the whole process of the interpretation of covert dogwhistles is not to be confounded with the recognition of speaker intention, as is the case with RSA models, because once the intention behind the covert dogwhistle is found, its effect supposedly disappears (Albertson, 2015; White, 2007). It is true that a lot of pragmatics (and especially game-theoretic pragmatics) is formulated in terms of retrieval of speaker intention. In contexts of information communication, where the CP is effective, *interpretation* can to some extent be reduced to the inference of speaker intention by the listener. It does not sound to me out-of-place that even though situations of political discussion are beyond the situations to which the CP applies, its core mechanisms necessarily become inefficient. An argument opposing the view in Saul (2018a) will be presented in section 4.2.

4.1.3.2 *In Henderson and McCready (2019b)*

The classification put forward in Henderson and McCready (2019b) takes another approach and classifies dogwhistles according to the nature of the content that they communicate, leaving us with two *types*:

1. *Type 1*: Type 1 dogwhistles are defined as dogwhistles that communicate social meaning in the form of the speaker being identified with a given group; *they do not communicate propositional content*. These are the verbal secret handshakes that we mentioned before, like an accent, but only heard by the ingroup. In Henderson and McCready (2019b), (16) is analyzed as being a type 1 dogwhistle.

2. *Type 2*: Type 2 dogwhistles are dogwhistles that *communicate propositional content*; phrases and words such that when they are recognized by the ingroup, trigger an interpretation of the utterance that differs from that of the outgroup. More specifically, the dogwhistles first communicate social meaning to the audience, like type 1 dogwhistles, and in virtue of the social meaning convyed, a propositional interpretation is triggered in the ingroup.

In Henderson and McCready (2019b), "inner cities", and specifically its use in (15a), is analyzed as a type 2 dogwhistle.

We can see already that the classifications differ in key aspects, and notably due to the fact that the approach in Henderson and McCready (2019b) is much more content-centric. We can say a few things about this classification. First of all, it does not seem clear that no propositional content is communicated in (16). The account in Saul (2018a) does mention that some content is communicated[17], and I would argue that even the glossing of the examples given in Henderson and McCready (2019b), already presented in (3) and reproduced below, underlines the existence of a communicated propositional content:

(17) a. Yet there's power, *wonder-working power*, in the goodness and idealism and faith of the American people.

b. Yet there's power, *Christian power*, in the goodness and idealism and faith of the American people.

Emphases mine

Maybe there are true type 1 dogwhistles, but this does not, to me, sound like a prototypical example of it. This is not a problem in and of itself for the account in Henderson and McCready (2019b), because the model that is proposed in the paper treats type 1 dogwhistles as a special case of the more general type 2 dogwhistles. We could still talk of the *type 1 content* of a dogwhistle to refer to its social meaning content, and of *type 2 content* to refer to its propositional content.

The fact that the differentiation is made on those specific examples however underlines the difference we'd already noticed between the two and highlighted by Mark Liberman's position[18], it could be that there is a difference, e. g.in the availability of plausible deniability between type 1 and type 2 contents.

4.2 DOGWHISTLES AND NON-COOPERATIVE COMMUNICATION

We've underlined already that standard, Gricean approaches to pragmatics are not necessarily appropriate to describe and discuss dogwhistles in virtue of the CP not being enforced in (notably) political speech. To this we can add that the account of implicit dogwhistles in Saul (2018a) does not, at first glance, sound compatible with the approaches we've seen so far of game-theoretic pragmatics, which rely "on the communication of a particular proposition" and "include the idea that recognition of the speaker's intention is required for success". This

17 About (16): "There are two messages a fundamentalist might take from this. The first is a kind of translation into their idiolect, to yield an explicitly Christian message that would alienate many [...] The second is simply the fact that Bush does speak their idiolect – indicating that he is one of them."

18 See footnote 14.

is in keeping with what we presented in Chapter 2: a lot of the work in pragmatics and interpretation has relied on the notion of cooperation, which is not to be assumed in the case of dogwhistle communication.

Saying that Gricean pragmatics is not appropriate to describe dogwhistles is not very controversial; typical Gricean pragmatics is not adequate for the description of many instances of natural language. It is for example hard to make sense of the voluntary flouting of, e. g., the maxim of quality, if we assume that the CP is enforced. And yet this specific maxim is flouted every time someone lies, and lying is not a particularly exotic thing to do. As was already discussed in Chapter 2, it is probably fair to say that Gricean pragmatics were developped with only a specific subset of language in mind and are therefore inadequate for the description of many things. Specifically, Gricean approaches are not necessarily adequate for the study of communication that is not content-based; most traditional approaches to both semantics and pragmatics have focused on the information about the world that is communicated in utterances, whether the information is directly derivable from the sentence itself or is inferred from the utterance and its context. In keeping with an earlier quote, pessimism about traditional approaches to meaning is explicitly mentioned in Saul (2018a):

> *"Fully making sense of politically manipulative speech will require a detailed engagement with certain forms of speech that function in a less conscious manner – with something other than semantically expressed or pragmatically conveyed content; and with effects of utterances that are their very point and that nonetheless vanish as soon as they are made explicit. None of the machinery developed in detail so far is equipped for this task."*
>
> — Saul (2018a)

In Saul (2018a), the concept of "covert perlocutionary speech acts" is presented as a way to characterize dogwhistles, and specifically implicit dogwhistles. Let us recall from Chapter 2 that *perlocutionary acts* is an expression typically referring to the effects of a locutionary act on the addressee. In Saul's theory, implicit dogwhistles communicate no particular content (compared with semantically equivalent expressions) and no particular social identity, they only exist through the effect that they have on listeners. The label *covert* perlocutionary speech act underlines the fact that in case the intention behind the dogwhistle is discovered by the listener, then the speech act fails. This approach allows her to avoid treating dogwhistled content as either propositional or intentional.

The notion of *intention*, and more specifically the need to go beyond it, is key in the theory presented in Saul (2018a). Similar positions are held in Beaver and Stanley (2018), where the intention of the speaker is seen as very secondary when trying to describe language beyond the subset consisting of communication of knowledge about the world. It is fair to assume that when Beaver and Stanley (2018) underline the need to escape from "ideal philosophy of language", they are to some extent referring to the fact that standard approaches to language,

especially in formal linguistics, do not necessarily seek to expand this subset of language deemed relevant for the study of meaning.

In Beaver and Stanley (2018), a proposition for theorizing the study of linguistic meaning based on political language is presented. The idea behind this is to focus on a subset of language that seems at first glance very different from the mere sharing of information about the world. In the context of political language, from public speeches to debates to pamphlet writing, while some information about the world is communicated, in a very "ideal language" manner, most of the time the goal is not truthful information sharing – or rather that's not the only goal, and the information sharing is sometimes more of a by-product of language use. The important bit of language in the realm of politics is rather the illocutionary and perlocutionary dimensions it has. It's not so much about what is *said*, but about what is *done*: criticizing, inciting, taking a stance, defining or showing off a personal brand... Studying political speech means focusing on these speech act intentions and their associated effects – on the illocutionary and perlocutionary acts – and not on the sharing of content or information. In fact, Beaver and Stanley (2018) underline more specifically the importance of the perlocutionary dimension, which was generally left out in earlier works in pragmatics, notably because of its *unintentional* aspect. Perlocutionary acts are beyond the intention of the speaker by definition, but they are central to the study of political speech and more generally speech that goes beyond the simple communication of information.

While the points of view presented in Beaver and Stanley (2018) and Saul (2018a) are certainly sensible (again, all the more so when discussing implicit dogwhistles), there is more to say about them:

1. While I do agree that the perlocutionary dimension of speech acts, and specifically of dogwhistles, is important, I disagree that those perlocutionary acts are to be seen as entirely outside the intentions of the speaker. In fact, I think that the notion of *covert perlocutionary speech act* put forward by Saul (2018a) does underline the importance of the speaker's intention.

2. Whether or not the communication of information is the main goal behind the use of language in a political setting, some information is indeed communicated in virtue of the fact that words have meaning; whether that meaning is intrinsic to the words and sentences or is negotiated at the scale of the exchange via pragmatic reasoning is irrelevant, there is no reason to think that the cognitive mechanisms behind meaning-making stop being effective when in a political context. So at least as a by-product, there is meaning in political utterances.

3. Because these cognitive mechanisms exist, it is likely that they are abused when needing to achieve goals that go beyond the mere communication of information. Political language is duplicitous, but it is still language in use, there is no reason to think it gets a free pass on some of the phenomena that we have otherwise observed. Moreover, politicians,

> and especially those partaking in any form of election, communicate about programs and issues and propose sets of measures to address these issues. The actual *content* of their discourse has some importance. While being duplicitous with regards to what is said is probably present, and in fact might be prevalent, voters do still rely at least in part on the content of the utterances to make their decisions, if only because they have no choice when wanting to impact the future of their country or preferred jurisdiction.

One grievance I will have about the picture put forward in Beaver and Stanley (2018) is the way they present dogwhistles, which in many ways is not unlike Saul (2018a). While I understand their focus on "plausible deniability", I will disagree with the following analogy:

> *"The whole point of words like "inner city" is that they are supposed to have a neutral explicit content. But it is not just dogwhistles that exemplify this kind of communicative plan. Take, for example, when someone says to someone else, "I'm not doing anything tonight." The speaker may mean to cause in their interlocutor a belief that the speaker is available for a date. But perhaps the speaker explicitly does not want their interlocutor to be able to calculate from general principles that they intended this. The speaker may speak in such a way as to mask that intention in order to avoid potential embarrassment. In such a situation, the speaker wishes to maintain plausible deniability."*
>
> — Beaver and Stanley (2018)

My reading of this analysis is that plausible deniability means that speakers want their intentions to be *vague* in a sense. I argue that they rather want their intentions to be *hidden*. The way dogwhistles have been described here, they are not *understood as* being ambiguous. In fact, that's the point behind using them, they are ambiguous in the sense that they can be interpreted in various ways, but that ambiguity is not acknowledged on the part of the listeners, otherwise listeners would understandably be dissatisfied when they are used. It is not the case, most of the dogwhistling in discourse is not noticed by most of the audience, which is the main reason why they are used. What I mean by a *hidden* intention that is not *vague* is that each listener, according to their own knowledge, group and/or ideology, will have their own interpretation of the dogwhistle that is being uttered[19]. Some will interpret that "inner cities" might refer to African-American neighborhoods (maybe exclusively), other will not, but there is no ambiguity there. The ambiguity is only seen by the people who are aware of the dogwhistle and call it out to underline the ambiguity, and by the speaker, who used the dogwhistle specifically because of this property. The ambiguity does exist, but the awareness of it is profoundly asymmetric, with only some people perceiving it. In short, the idea behind dogwhistles as they are understood here is that their conventionalization is differentiated according to groups[20]

19 I do not wish to say that the interpretation is personal in the sense that it is not socially determined, I mean to say that the interpretation mechanism is individual-based, even if it obeys to rules that emerge at the level of the group.

20 For a similar approach to the question, with a system of dogwhistle types that relies more on traditional pragmatics approaches, see Guercio and Caso (2021).

One reason I think this is that from a speaker point of view, choosing utterances that are more ambiguous to everyone is not necessarily a good strategy. If the in-group, for example, sees an ambiguity, then they might make the inference that the speaker is dogwhistling, but crucially, and as underlined in Saul (2018a), that is not good, because if the listeners of the in-group realize that there is political manipulation happening, they might not know from which side it comes. Speakers do not want the in-group to notice the dogwhistle for what it is, they want the in-group to notice their belonging to their group and shared ideology. If we go back to the account of the use of religious cues given in Kuo (2006), Bill Clinton's use of them was not effective perhaps because it was too obvious that it was political manipulation, based on his general ideology; George W. Bush's use of them was more effective because his general persona and ideology are more Evangelical-compatible, yet it was also political manipulation to the detriment of Evangelicals, as underlined by Kuo's conclusion. Because the political manipulation can work both ways, it is in fact not good that the intentions of the speaker be ambiguous to the audience, including the in-group.

My position on this is that dogwhistles, both explicit and implicit, do rely on a mechanism of interpretation of the intention of the speaker, but that there is a sizable difference (in fact a type difference) between the intentions inferred by the listeners and the actual intentions of the speakers. Dogwhistles seem to me a means to communicate many things at once to people who would typically infer only one of the things; it relies on a habit of using language univocally and abuses it for multi-vocal purposes.

In Beaver and Stanley (2018), the idealizations about language that are usually done in the study of its meaning are said to "force us to restrict our attention to cases in which we either literally express the belief we want our audience to adopt or communicate that belief in a way that allws them to calculate our communicative intentions [...] while assuming that we are inviting them to make this calculation and attribute these intentions to us". To me, dogwhistles are not that far from this; they communicate *a belief* in a way that allows listeners to calculate that belief and attribute it to the speaker. The difference is in the indefinite. The authors later claim that "the paradigm of a failure to use language correctly, according to the standard theory of meaning, is to misrepresent the world, to lie". Seeing lies as a failure to use language correctly is indeed a mistake, but seeing it as a failure to *cooperate*, in favor of principles that, perhaps, supersede that of cooperation, seems a lot more reasonable[21]. Dogwhistles, like lies, are not cases of cooperative communication. Like lies too, their utterance can nonetheless be understood as such by unsuspecting listeners. The real need that there is is not simply to get rid of the notion of cooperative communication to describe them, but to acknowledge that communicative principles might not be shared by speakers and listeners.

21 I do not think anyone truly holds the view that *lying* is a *failure to use language correctly*, this reading stems from too strict a reading of Grice (1967). The CP as initially thought of is not, in fact, the be-all and end-all of language use. A better reading of Grice is more simply that there are some principles that govern language use, and inferences from signals can be made thanks to those principles. It is not incompatible with the idea that there could be several principles beyond the CP, with some perhaps superseding it, and that superseding being context-dependent. It is however the case that these cases have not been explored too much in the formal semantics and pragamtics literature.

4.2.1 *Concluding remarks*

This concludes a first approach to the concept of dogwhistles. Crucially, we have not tackled the issue of how they emerge and how they evolve, this will be explored a bit in Chapter 9, where we will present some theories about the appearance of dogwhistles as well as some regarding their long-term effects on discourse.

For the moment, however, we have seen what dogwhistles are with a few examples, why they are used, and have discussed some of the categorizations present in the literature. The rest of the work will focus in particular on explicit intentional dogwhistles, but at stated in 4.2, the distinction between explicit and implicit dogwhistles, in terms of mechanisms, is not necessarily as clear-cut as is presented in Saul (2018a). Specifically, I have argued that concepts generally found in the study of meaning and thought to be inadequate in Beaver and Stanley (2018) and Saul (2018a) are not necessarily as out-of-touch with the concept as is defended in those works.

One thing is however certain: dogwhistles appear in contexts that are very different from the usual objects of study of pragmatics and semantics, where a principle like the CP does not necessarily apply, which will require some deep revisions regarding existing systems, like RSA and SMG if we want to adapt them to their study.

When discussing the study of non-ideal language and its formalization, four possibilities are put forward in Beaver and Stanley (2018):

> *"The first option is that extending the scope of the theory of meaning may require no adjustment to the formal structures and tools or their interpretation; it will involve at most adding some new tools. The second option is that extending the scope of the theory of meaning may necessitate a new understanding of the formal structures and tools, but it will not require us to abandon them altogether. The third option is that extending the scope of the theory of meaning will lead to a dramatic reformulation and reimagining of the theory of meaning, so that previous work will have to be discarded. Finally, the fourth option is that extending the scope of anything like a theory of meaning to the phenomena we have discussed is impossible, that these phenomena are resistant to analysis by anything with the structure of a formal theory. If this fourth option is correct, then a theory of meaning that includes political speech is not possible. All that is possible are particular descriptions of practices in their historical contexts. "*
>
> — Beaver and Stanley (2018)

This work, and especially what is presented the next two chapters, will adopt the position of formal optimism and assume that the options 1 and 2 are valid. I am very afraid of the other options.

5

A FORMAL APPROACH

"Meaning is born out of struggle: the struggle in the (partly conscious) mind of the speaker to mark, among the available morphological counters, the one most able to articulate the inarticulate bubble of possibility the unconscious proposes."

— T. Price Caldwell, making a case for the emergence of meaning through pragmatic reasoning rather than intrinsic semantics. (Caldwell, 2018)

"If a book were written all in numbers, it would be true. It would be just. Nothing said in words ever came out quite even. Things in words got twisted and ran together, instead of staying straight and fitting together. But underneath the words, at the centre, like the centre of the Square, it all came out even. Everything could change yet nothing would be lost. If you saw the numbers you could see that, the balance, the pattern. You saw the foundations of the world. And they were solid."

— Shevek, physicist, being extremely optimistic about the translation of natural language into mathematical concepts when thinking about magic squares. (K. Le Guin, 1974/2019)

The work presented here is based on and expands upon Dénigot and Burnett (2020)

We have seen in Chapter 4 that the term "dogwhistle" can refer to a number of phenomena that are in fact quite different one from the other. There is specifically a huge discursive difference between *explicit* dogwhistles, which are closer to the communication of a secret message to a group of people (either in terms of communicating propositional or socially meaningful information) and *implicit* dogwhistles, where speakers, intentionally or not, trigger specific thought processes in their audiences. The first group is closer to sending a *coded message* – with the key difference here that the cipher used is partly accessible to any eavesdropper – while the second would be closer to a subliminal message – in the sense that it triggers unconscious thought processes.

This part focuses on the project of giving a formal description of dogwhistle communication, using the concepts of game-theoretic pragmatics and SMG that we have presented in Chapter 2 and Chapter 3. We will attempt to translate the key properties of dogwhistles in the language of game-theoretic pragmatics and see what kind of formal object we end up with. As stated before, we will mostly focus on *explicit* dogwhistles, since at first glance it seems that the game-theoretic pragmatics toolbox is more adapted to describing those, but in the spirit of Beaver and Stanley (2018), we will also discuss the extent to which these considerations can also apply to cases where the notion of *intention* is not as seminal.

5.1 REEVALUATING SIGNALING GAMES

We have seen in Chapter 2 and Chapter 3 that GT provided us with *signaling games*, a class of games that, following Lewis (1969), has been widely used to try and describe the exchange and interpretation of linguistic signals, leading notably to the RSA and SMG frameworks. In the case of signaling games and of RSA-based models, we have underlined the importance of a certain degree of cooperation, the goal of the game being, after all, the reliable sending of information. Yet in Chapter 4, we have seen arguments supporting the idea that situations of political discourse, and specifically situations where dogwhistle communication arises, are not typically focused on the notion of cooperation. Given that speakers make an effort to veil their intentions to at least part of their audience, we are here in a *non*-cooperative setting.

A question arises then: can game-theoretic models based on a tradition of truthful sending of information be adapted to describe non-cooperative situations like dogwhistle communication? In the context of this thesis, as stated in section 4.2.1, we assume that yes, RSA models can serve as a basis for the description of dogwhistle communication, but it goes without saying that some alterations will have to be brought in.

The model presented in the next sections will be based upon both RSA and SMG, but contrary to both of those models, we will have to assume here that *there are several listeners*. Not only that, but there are also several listener *types*. The concept of a many-faceted crowd is crucial in discussing dogwhistles, since they are conceived to satisfy people with differing opinions.

In game-theoretic terms, and specifically in Bayesian games, a player's *type* is usually understood to be its utility function, which in turn reflects each player's preferences. Remember that in a standard signaling game, the emitter of the signal's type is understood to be the state of nature. Translating that into the specific RSA interpretation of signaling games, a Speaker's type is the information that they have observed (and therefore wish to communicate). There are no Listener types, however, because all Listeners have a single goal, which is to have a correct interpretation of signals. In our dogwhistle scenario, this is still thought to be the case, but we will define Listener types in the sense that different Listeners have different prior beliefs about the Speaker, and have access to different *lexica*. In addition to Listener types, we will present different Speaker *families*. Those families are not *types* in the sense that we have just described here, nor are they Speaker *types* in the sense of classical signaling games, they actually constitute a variety of solution concepts for speakers to the game that we will describe. More on both Listener types and Speaker families in Section 5.3.1.2.

Lexica are a key point of our model. We can find a notion of *lexica* presented in Bergen, Levy, and Goodman (2016) for the description of M-implicatures[1]. We have here something similar. The key idea is that words of a lexicon can be interpreted in possibly vastly different ways. In

1 M-implicatures are implicatures that are thought to arise when "nonstandard" formulations are used. More specifically, the idea is that more complex utterances (in the sense, for example, that they use more words, or in general more convoluted ways to express an idea), when compared with simpler, but truth-conditionally equivalent utterances, will trigger implicatures of non-prototypicality. So if I say "Tracy killed Taylor", the listener makes no particular implcature, but if I say "Tracy caused Taylor to die.", I might make the implicature that Tracy killed

the case of Bergen, Levy, and Goodman (2016), they use this to add uncertainty to the Speaker of their model, this added uncertainty allows them to properly derive M-implicature effects. In our case, we are not adding uncertainty to the Speaker, the lexica are here to symbolize the idea that people from different social groups use words in a different way, possibly with different propositional meanings. In a way, one could think of the interpretation of dogwhistles that we have in our model as an issue of overlapping dialects. This interpretation seems to be the most natural in the sense that it allows to explain away the fact that socially meaningful information can be directly linked to difference in interpretation. It does not come without issues however.

This interpretation might not be satisfactory in the sense that it would not allow in itself to explain why the dogwhistle terms are not used within the targeted ingroup when there is no member of the outgroup present (or at least we don't suppose they are), since it is part of their dialect after all. This limitation is mentioned in Henderson and McCready (2018). We argue here that this can be explained using standard RSA and scalar implicature phenomena, whereby more explicit/less ambiguous lexical items would be preferred; but even then, we partly disagree with the premise, it does not in fact seem far-fetched to say that, e. g., evangelicals use the same kind of terminology among themselves and in speeches. It is the main reason why the social identification of the speaker as an evangelical works.

Another limitation of this approach lies in the fact that it treats the dogwhistled content as at-issue; while this is indeed problematic when talking about implicit dogwhistles (at least in the interpretation we have seen in the previous chapter), we argue that it is not as much of a problem when talking about explicit dogwhistles, which goes to underline strong differences between the two. From a philosophical perspective, we adopt the point of view that actually, explicit and implicit dogwhistles are fairly differents beasts, and that it may not necessarily be adequate to treat them as two instances of the same phenomenon. In spite of this, following, e. g., Henderson and McCready (2018), we still focus here on the "inner cities" example, both to facilitate the comparison between our approaches and because the "inner cities" example provides us with a much easier analysis in case we want to study the possible conveying of supplementary semantics than the "wonder-working power" example.

An added value of the overlapping dialects approach that we propose is that it easily deals with the issue of *synonyms*. If there is no added propositional meaning to dogwhistles themselves and they only carry stereotypes and trigger specific thought processes, how come we do not observe dogwhistle effects with at first glance truth-conditionally equivalent lexical items? The dogwhistle effect seems to be linked to specific expressions, and any formalization of the phenomenon should take this into account. The overlapping dialects approach allows us to easily evacuate the issue.

Finally, because we make the assumption that both social meaning and content are communicated in dogwhistles, our approach will try to merge standard RSA, which focuses on

Taylor in a non-prototypical way. The intuition behind it is that the added cost of the complexity of the sentence is used to convey meaning.

implicatures and content, and SMG, which uses similar foundations to discuss the communication of social meaning.

5.2 PREVIOUS APPROACHES

5.2.1 *Various inspirations*

Before tackling the previous attempts at a formalization of dogwhistles *per se*, we will first focus on formal approaches to other linguistic/discourse phenomena in which we have found inspiration.

As mentioned above, we present a view of communication that revolves around the idea that different lexica may coexist in a single speaker/listener, a concept that is present in Bergen, Levy, and Goodman (2016). Specifically, staying in a RSA framework, what we find in Bergen, Levy, and Goodman (2016) is uncertainty regarding the interpretation of messages by the Literal Listener. Remember from Section 2.3.2 that the Literal Listener is the purely theoretical representation of a non-pragmatic, unrefined listener that serves as a starting point for the recursive reasoning in RSA that leads to the emergence of implicatures. The idea in standard RSA is that such a Listener has a straigthforward, purely semantic interpretation of utterances (and is therefore unable to decide between two interpretations in case an utterance can semantically refer to both). In Bergen, Levy, and Goodman (2016), we are also faced with the fact that there may be a variety of different, possibly opposed semantics (or *lexica*) for any given utterance, with a Literal Listener being equally likely to use any of those. The concept of *lexica* that we use here functions similarly in the sense that it consists in a variety of possible semantics accessible to Literal Listeners, but it is different in the sense that each lexicon is considered to be fixed and chosen a priori. This will be made clearer in Section 5.3.1.

Another body of work in which we can find some inspiration is that presented in e. g.Asher, Paul, and Venant (2017) and its framework of Message Exchange Game (MEG). The main insight that we will take from this work is the notion of "jury", and the fact that it is typed. In a few words, the "jury" of a MEG is a (possibly fictional) player that will decide who "wins" an exchange, in the sense of whether speakers attain their conversational goals, including in adversarial contexts. It can be seen as a way to enforce social norms in discourse. While we will not have a "jury" *per se*, we will have typed listeners belonging to opposing groups, with one of these groups enforcing a notion of social norm. Intuitively this is what happens with dogwhistles, both explicit and implicit, where the so-called outgroup acts as a guardian of norm. If a norm like the "norm of racial equality" from Mendelberg (2001) is overtly breached in discourse, then the outgroup of the dogwhistle will punish the speaker in terms of reduced support. We will expand a little bit on the MEG framework in Section 5.2.3.

This leads us to one of the key differences between standard signaling games and our model: the Speaker conveys a pîece of information, but has a choice over that information.

In political speech, again, the goal is not the truthful sending of information, it is rather personal branding, story-telling, or support gathering. Some content is communicated, but that content is not dictated by anything other than the personal preferences in terms of content communication of the speaker. As was stated in Chapter 4, dogwhistles, both explicit and implicit, are duplicitous in nature, we interpret that duplicity as meaning that speakers exploit usual patterns of thinking based on cooperation to achieve their goals. This is why we still use frameworks like RSA, which were conceived with cooperation in mind: our position is that the dogwhistle effect arises because (some) Listeners assume that the Speaker is being sincere enough that they can infer from their speech the implicatures they would usually derive, and the Speakers play on those expectations. In this respect and others, we are closer to the SMG framework, where the notion of Speaker preference is much more significant.

5.2.2 *The Henderson & McCready approach*

The most prominent formal rendition of dogwhistle communication is found in Henderson and McCready (2018, 2019a,b). We will focus here on the presentation given in Henderson and McCready (to appear), which is the most up-to-date[2]. The approach put forward in Henderson and McCready (to appear) relies on inference from social meaning communication. In short, dogwhistles in that theory are first and foremost tools used to communicate social meaning to the audience; in that respect, they do act like dialectal differences in the sociolinguistics sense. Whatever propositional content (if any) is communicated through dogwhistling is however not understood in a dialectal way, but is entirely due to listener interpretation. That is, there is no meaning conveyed by the dogwhistle other than "I am part of the in-group". Members of the in-group may or may not then make supplementary inferences regarding propositional content.

More precisely, the models presented in Henderson and McCready (to appear) focus on the tension regarding conventionalization in dogwhistles, specifically between the accounts of Stanley (2015) and Khoo (2017). In Stanley (2015), we are presented with an account of dogwhistles as conventional implicatures, with dogwhistles communicating non-at-issue (NAI) content as well as at-issue (AI) content. This was hinted at in Chapter 2, we will reproduce here (9) for explanatory purposes.

(18) Fargoth is such a n'wah[3].
  ⇝ Fargoth is a foreigner. AI content
  ⇝ The speaker dislikes foreigners. NAI content

2 It also happens to be fairly different from the previous accounts in Henderson and McCready (2018, 2019b), confronting more directly the contribution of Khoo (2017). Many thanks to Elin and Robert for allowing me to look at their draft.

3 See footnote 17 in Chapter 2.

(18) leads to two implicatures, one of which is AI, the other NAI content. The idea is that NAI content is not a proper part of the semantics of an expression, it is a dimension of meaning that the expression has acquired through convention of usage. *N'wah* literally means *foreigner*, it just so happens that it has been used to refer to them in a derogatory way. The conventional, NAI meaning at play here is an attitude of the speaker regarding an issue. But because it is conventionalized, it will sound very strange to answer something like (19c) to (19b):

(19) a. Fargoth is such a n'wah.

b. What do you have against foreigners?

c. ?I have nothing against foreigners, I'm just stating a fact.

In other words: there is little deniability to conventional implicatures. Compare with (15a). If Ryan were to be called out on the dogwhistled content in his affirmation (which he was), he could without any obvious problem answer something like (20c) (which he did):

(20) a. We have got this tailspin of culture, in our *inner cities* in particular, of men not working and just generations of men not even thinking about working or learning the value and the culture of work.

b. Are you saying African-Americans are lazy?

c. I have never said that, I don't have a racist bone in my body.

This illustrates the idea that the content of dogwhistles is not (fully) conventionalized. If it were, (20c) should sound out of place, which it does not. This is notably underlined in Khoo (2017), who proposes that there is no such thing as dogwhistled *content*, that the use of dogwhistles merely allows for subsequent inferences on the part of the listeners, making them fully deniable. In short, if a listener holds a certain belief about something, then mentioning that thing will lead to that belief being effective. In the case of (15), if a listener has the belief that most inner cities in America are in fact African-American neighborhoods, then hearing the term "inner city" will lead to them assuming that we are talking about "African American neighborhoods".

But as underlined in Henderson and McCready (2019b, to appear) this account is also lacking in the sense that it cannot explain why expressions with identical denotational semantics to dogwhistles do not trigger those inferences. If we conventionalize dogwhistles, then they become undeniable; if we have no conventionalization, then we fail to capture the difference between "inner city" and "city center". The middleground solution to that conundrum proposed in Henderson and McCready (to appear) revolves around the idea that dogwhistles are conventionalized from a social meaning point of view, but that they don't convey conventionalized truth-conditional content. The resulting model predicts that upon hearing a dogwhistle, members of the ingroup will recognize the speaker as one of their own and will link that speaker's persona to the *ideology* of the group, ideology being here construed as a set

of assumptions about the world and stances regarding parts of the world or other groups of people, a shared *common ground* of sorts.

This approach allows to describe most relevant processes involving dogwhistles, but there are a few points in what they present that we could try to address:

First of all, the model that they propose is based on SMG, but with an important addition: whereas SMG treats social meaning as an absolute, much in the way of meaning in an extensional semantics sense, the models in Henderson and McCready (to appear) do not do that and have a prior distribution over the social meanings of given utterances. While this is philosophically justifiable (social meaning is not propositional in nature anyway), it does beg the question of knowing where these priors come from. A key insight behind models like RSA is that you do not need to start from probabilistic semantics to end up with probabilistic interpretations, and that probabilistic behavior can emerge from boolean premises. The Henderson and McCready (to appear) approach loses this on the social meaning front.

There is a strong insistence on the fact that there is no propositional content to dogwhistles, only socially meaningful content, but while the ideology approach sounds very promising, it calls for a number of formal objects that are very hard to define (including, above all, ideologies themselves), and in the end requires a full working model of human societies at large to fully account for the triggering of new propositional content in dogwhistle understanding. I am not saying that it is not necessary, in fact I think that this is also necessary to give a full-fledged rendition of social meaning in SMG frameworks, but I think that positing a degree of conventionalized propositional content to dogwhistle terms can make the process a lot simpler and still account for a number of phenomena, if not all of them.

The model as it is presented relies on the fact that whenever a speaker's persona is recognized, it will be linked to a corresponding ideology, and the triggered inferences are then the effect of the underlying ideology. Roughly, what happens is speaker S uses m, if listener L is part of a given ingroup, they recognize m as indexing that ingroup, and any supplementary meaning associated with m comes from uncovering S's group and its ideology. But what if S's group is transparent? Is the dogwhistled propositional content necessarily uncovered by listeners? The Henderson and McCready (to appear) approach states that as long as you have knowledge of the association between personae and ideologies, and have knowledge of the tenets of said ideologies, then detecting the dogwhistle allows you to infer the dogwhistled content. This works well with the idea that once people are made aware of the dogwhistle, they might disapprove of it (Albertson, 2015), but it does not explain scenarios like the one presented in the Griffin quote in Chapter 4, where the persona is fully known by the audience (as is, in fact, often the case for political discourse), but where using dogwhistles is still accepted by the audience and their dogwhistled content still evades the audience. In Albertson (2015), while it is true that only members of the ingroup truly recognize associated personae when dogwhistles are used, what triggers the rejection of discourses by outgroup members is not the revelation of the identity of the speaker, it is the enriched meaning of the utterance itself.

In the case of the "wonder-working" example in (16), people would not reject the evangelical persona. There is no shame in being an evangelical in mid-2000s USA. What they do reject is the presence of overtly religious speech in politics. If they start rejecting the dogwhistle once they are aware of its origin/meaning, it is because they are rejecting it on the grounds of it having become an overtly religious appeal, when before that it was just conservative discourse. It sounds as if the ideology-based account proposed in Henderson and McCready (to appear) leads to people outright rejecting entire characters on the basis of their ideology, but is this what happens or do they reject, rather, the statements themselves? Is there no case where known far right figures using polished language are commented upon as "making sense" or "asking the right questions", when the underlying interpretation within their community is in fact miles away from the overt content of what they are saying?

This is an important part of dogwhistle strategies: making specific ideologies palatable through ambiguous statements so that people accepting the statements themselves, might gradually accept the ideology; or rather so that a new ideology might be constructed based on those dogwhistles. Ideological content can take a propositional form, in that case, what if some ideologies are based on statements that are themselves dogwhistles? What would then be the content of that ideology? Look for example at the recent leadership crisis in the USA-based far right group Proud Boys. While the organization has repeatedly insisted that it did not embrace racist beliefs and existed for the defense of "Western civilization", it appears that for at least one subgroup in the organization, the terms "West" or "Western civilization" are code for "white race"[4]. This can come as a surprise, especially given the fact that since 2018 (and at the time of writing), the chairman of the organization does not identify as white but as Latino. A Henderson and McCready (to appear) account of this situation would be that although Proud Boys' members' ideologies are similar, they are not identical and notably differ on the race front. Another reading would be that initially overtly racist members have pushed forward a watered-down version of their ideology in order to seduce more people into the movement, effectively hiding part of the ideology/constructing the ideology specifically by using dogwhistle statements like "Western civilization".

Following Khoo (2017), Henderson and McCready (to appear) insists on the fact that the dogwhistle statements themselves have no dogwhistled propositional content; what I am arguing for is that they do, at least to some extent, and that when people oppose dogwhistle statements on the basis of the underlying ideology, they do not oppose the entire ideology, but only specific positions from those ideologies. It would be silly to have an exchange like:

(21) a. We have got this tailspin of culture, in our *inner cities* in particular, of men not working and just generations of men not even thinking about working or learning the value and the culture of work.

b. ?Are you saying non-whites are replacing whites?

4 As reported, e. g.here: https://www.adl.org/blog/proud-boys-bigotry-is-on-full-display.

It would be silly because (21a) does not say anything about a change of demographics, even though beliefs like the one alluded to in (21b) are beliefs usually associated with white supremacist ideologies. If the part of the ideology that is being hinted at by the use of a specific dogwhistle is stable, then there is some degree of conventionalized propositional content. What I argue for in the next sections, when presenting my own formal model, is that dogwhistles do have conventionalized propositional content, and that that propositional content is directly communicated to the ingroup.

5.2.3 *Message Exchange Games (Asher, Paul, and Venant, 2017)*

If we are to talk about non-cooperative approaches to discourse, MEG (Asher and Paul, 2018; Asher, Paul, and Hunter, 2021; Asher, Paul, and Venant, 2017) is an important framework in the domain that needs to be expanded upon, even though we will not directly base our formal modeling on it. One key insight of that framework lies in two main ideas:

1. Speakers are not necessarily cooperative
2. Language is a tool used to achieve specific goals beyond language itself

These considerations lead the research presented in Asher, Paul, and Venant (2017) to distance itself from the original concept of signaling games, which are intrinsically connected to ideas of cooperation and are generally useful for one-off interactions and describing situations where the understanding of language in itself and for itself are at stake – for example through a focus on specific linguistic phenomena like the interpretation of scalar implicatures. A very important part of MEG that contrasts them with standard signaling games, both from a philosophical and from a formal point of view is that they are ultimately *not* games that describe the interpretation of messages, but games that describe the *unfolding of conversation*.

This means notably that they are not limited to single turns like signaling games typically are, and are *infinitary*, meaning that even though any given conversation is finite, the games themselves are conceived as infinite, a characteristic that articulates well with a notion of non-Gricean speakers, since in the absence of cooperation and with conversation and dialogue becoming verbal jousting aimed at the achievement of possibly opposite goals, one cannot be thought of as knowing in advance the issue of a conversation and *who has the last word*.

For both philosophical and formal reasons, the winner of such jousts has to be determined by an entity outside from it, the *Jury*, a fictional player that may or may not be present and may or may not take part in the game. In some of the examples put forward in Asher, Paul, and Venant (2017), the Jury is very much a Jury, in a legal setting, with instances of individuals taking the stand and answering attorneys' examinations. It can also be thought of as being the general audience in a political debate, or even be something as abstract as a norm of politeness or political correctness. The point is that the Jury is a normative entity which determines who wins and achieves their conversational goals, leading to a situation very different from the

signaling games approach that we have seen before, where speakers do not address each other as much as they address the Jury and try to satisfy the expectations of the Jury. In a way, the outgroup listener in our model – along with the preferences of the speaker – play the role of the Jury, as it is their interpretation of the model that dictates whether the speaker will maximize their utility or not. We chose to rely on an existing listener to play the role of a normative instance[5], but there is a way of interpreting our outgroup listener as being philosophically akin to a Jury in MEG.

MEG is a very interesting and singular framework, and along with a proper representation of discourse parts (which it finds in Segmented Discourse Representation Theory, SDRT, Asher, 1993; Asher and Lascarides, 2003; Asher and Vieu, 2005), it is a powerful tool to describe the unfolding of dialogue, notably in non-cooperative contexts. Because it focuses on non-linguistic goals, it is also likely that it could easily be generalized to non-linguistic communication practices, like gesturing or calculated silence. So why wouldn't we use it?

In spite of all its qualities and insights, we probably do not need the heavy machinery of MEG to analyze cases that are ultimately, in our case, one-off interactions. While there are possible answers to dogwhistles, and while they can very much be used in debates, the cases we are interested in here are mostly unilateral communication between a speaker and an audience, with the audience not being in position to necessarily answer the speaker. Think, for example, about the *wonder-working* example from earlier. There is no dialogue there, and we need not see it as a dialogue. What there is is a single speaker, and very many listeners, all having their own interpretation of the signals they perceive. Besides, and even though standard MEG (Asher, Paul, and Venant, 2017) have been extended with notions of belief uncertainty (Asher and Paul, 2018) and bias (Asher, Paul, and Hunter, 2021), the entire system put forward in Asher, Paul, and Venant (2017) is not particularly focused on either social meaning or ambiguity resolution, and does not, in itself, allow for the use of lexical differences between speakers, which are all key points of our approach here.

5.3 DOGWHISTLE GAMES

As said above, the model presented here relies on the idea of different listeners using different lexica with overlapping semantics, and a speaker choosing to maximize the ambiguity when facing a diverse crowd. Before defining the model itself, we have to be clear about what properties this model should have:

- INTERPRETATIVE VARIABILITY: it has to be the case that interpretation of a dogwhistle differs from listener to listener.

5 This relies on the intuition that discourse normativity is most powerful when in the presence of an outgroup member, or at least a listener that is not considered an equal, whether because they are a stranger or above the speaker in a given hierarchy.

- POLITICAL CONFLICT/ADVERSARIAL CONTEXT: it has to be the case that using a dogwhistle is preferred when there is ideological conflict in the audience when compared with when there is no conflict.
- PLAUSIBLE DENIABILITY: it has to be the case that the dogwhistled content is deniable. This translates into the probabilistic interpretation of the dogwhistle never being equal to 1.
- CAGEY LISTENER[6]: our model has to account for listeners who recognize the dogwhistle while not being part of the ingroup. They are understood as being a special type of listener.
- FORM-BASED: our model should take into account that dogwhistles are specific linguistic items, and not just messages.

In addition to this, the model presented here is IDENTITY-BASED, meaning that the interpretative variability aforementioned is linked to the identites of the speaker and the listener.

In addition to this again, this model can be understood as the merging of standard RSA models for the derivation of scalar implicatures with SMG models for the communication of social meaning. We will explore this further in the rest of the chapter, but as is, the model can*not* reproduce the behaviors of either initial model. With a slight tweak, however, it can easily become a generalization of both.

5.3.1 *The Horrible Formal Model*

Given these considerations, we can define what a DWG is.

Definition 5.3.1. Dogwhistle game

A Dogwhistle game (DWG) is a tuple of the form:

$$\langle \{S, \{L_i, L_j\}\}, L_0, W, M, \Pi, \textsc{Lex}, \textsc{Soc}, \Pi\text{-}\textsc{Lex}_{i/j}, \Delta\text{-}\textsc{Soc}_{i/j}, Pr_W, Pr_\Pi, \alpha, \alpha', \beta, \beta' \rangle$$

In this tuple, we find the following:

1. S is the speaker[7].
2. L_i, L_j are two listeners[8].

6 In Dénigot and Burnett (2020), this is called a *savvy listener*, due to terminological overlap and to avoid confusion with the concept of savvy listener in the models put forward in Henderson and McCready (to appear), we have changed it here to this denomination.

7 We will present several types of speakers in the remainder of the chapter, but the most important one is the "duplicitous speaker".

8 Making these two listeners identical trivializes the game and makes using dogwhistles (mostly) useless.

3. L_0 is the Literal Listener. We can essentially define it as a function, like what we did for RSA in Section 2.3.2, but this time with two outputs: $M \rightarrow \Delta(W), \Delta(\Pi)$. Because it is a great deal different from the standard RSA L_0, it is defined in more detail in Section 5.3.1.1 and visualized in Figure 2.

4. W is a set of worlds w.

5. M is a set of messages m.

6. Π is a set of personae π.

7. Lex is a set of *lexica*, or interpretation functions $[\![\cdot]\!]$.

8. Soc is a set of *social lexica*, or indexation relations $[\cdot]$, as per Burnett (2019).

9. Π-Lex : $\Pi \mapsto \Delta$Lex is the socially-determined lexicon function, which maps specific personae π to probability distributions over available lexica in Lex. There is one such function per listener.

10. Δ-Soc is the set of priors over indexation relations $[\cdot] \in$ Soc there is one such prior distribution per listener.

11. Pr_W is a set of probability distributions over W, representing listener prior beliefs regarding the state of the world.

12. Pr_Π is a set probability distribution over Π, representing listener prior beliefs regarding the persona of the speaker.

13. $\alpha, \alpha', \beta, \beta'$ are all temperature parameters. Like temperature parameters used in standard RSA, they are used to make the choice of utterances and interpretations more or less flexible, with a higher parameter usually leading to less uncertainty. In the present work, although we have made them part of the model to respect conventions, we will mostly not use them in the examples that follow, setting all such parameters to 1.

As mentioned above, one key addition of this model when compared with the approaches in Henderson and McCready (to appear) is the fact that there are multiple interpretation functions available, reflecting the variation in lexicon use that can be found in different groups. The Π-Lex function links these interpretation functions to speaker personae. Depending on who one thinks the speaker is (in terms of social group), one will make inferences about how they are likely to speak. In other words: a single message can have different meanings depending on who says it (to whom). As with all RSA models, listeners derive their interpretation from what a speaker faced with a literal listener would say, but they also add considerations about the possible personae of the speaker (reflected in the priors they have about that aspect of the message). The strength of the listeners' assumptions depends on the speaker.

This model can be seen as both a listener model and a speaker model. As is customary, we will first describe the behavior of the literal listener.

5.3.1.1 *Literal Listener*

Let's start with the Literal Listener model. Although the spirit of this is very close to what is found in RSA models, the fact that there is a lot more information that is taken into account by listeners to interpret messages and by speakers to choose them leads to a more complicated picture than is usually found in standard RSA, notably at the level of the Literal Listener. Specifically, because we assume that the derivation of semantic meaning is dependent upon the social identity of the speaker, the choice of interpretation is done in several steps:

1. Reception of the utterance.
2. Updating priors on the speaker's persona from the utterance.
3. Choice of interpretation function based on the information from step 2.
4. Derivation of the meaning of the message based on the choice from step 3.

This imposes a specific sequentiality on the entire process. This is not something that we would necessarily want, as it is very likely that this process works both ways (meaning that the content of the discourse can influence a listener's perception of the speaker's persona). This issue is not addressed here, but we present a way to circumvent it, at least in spirit. Figure 2 presents the way the Literal Listener interpretation is derived. Using the values set a priori in Pr_W, Pr_Π, Δ-Soc and Π-Lex, we first derive the two *literal* meanings of the utterance: semantic and social, in the RSA and SMG fashion. After that step, in order, we have:

1. $P(\pi|m)$, the interpretation of a persona given a message, made using the literal social meaning of m and normalizing over all the possible social indexation functions $[\cdot] \in Soc$.
2. $P([\![\cdot]\!]|m)$, the estimation of the appropriate interpretation function upon hearing m, computed using the values from Π-Lex, $P(\pi|m)$ and taking all $\pi \in \Pi$ into account.
3. $P_{L_0}(w|m)$, the content interpretation, based on the semantic interpretation and the estimation of the most appropriate interpretation function.

The two output values, $P(\pi|m)$ and $P_{L_0}(w|m)$, are the ones we are really interested in in the end, because those are the ones that will be used by the Speaker for their own choice among messages. Beyond this step is where our model vastly differs from other implementations of RSA models and where we introduce a notion of *player families* that goes beyond the usual type-distinction usually found in such game-theoretic models (see Appendix A).

5.3.1.2 *Player families*

Starting from the Literal Listener just mentioned, we derive the following player families:

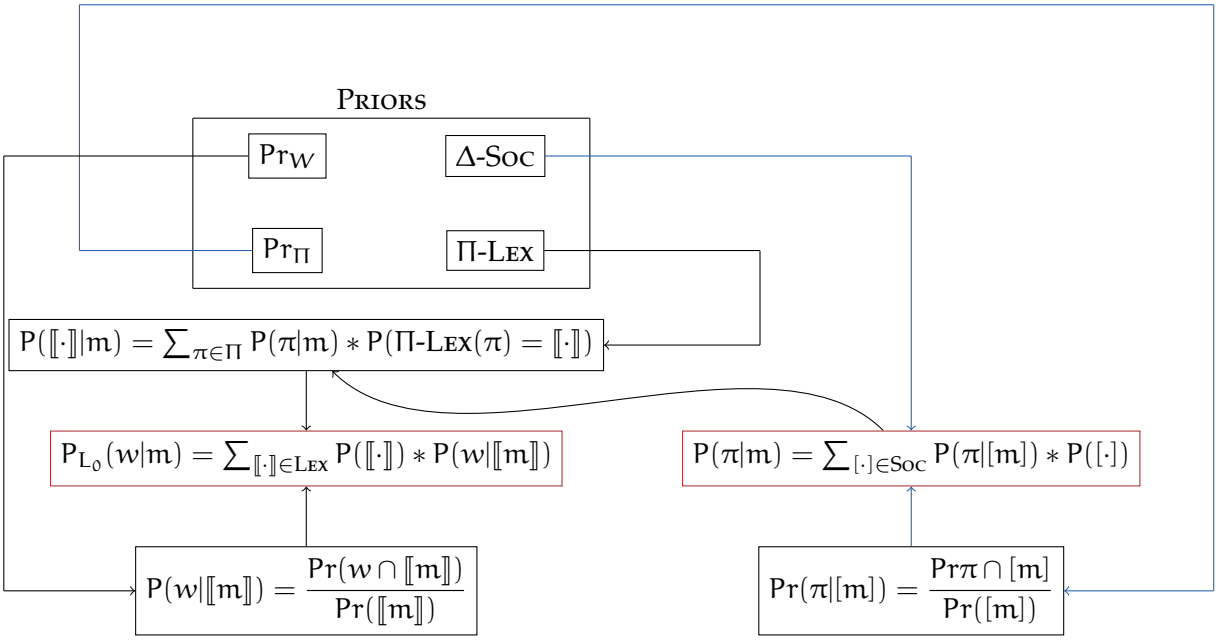


Figure 2: Overall view of the Literal Listener solution concept in DWG. The red boxes indicate the output computations that will be used by our Speakers. Blue arrows indicate computations related to social meaning, black arrows indicate computations related to semantic meaning.

SPEAKER FAMILIES:

1. *The cooperative, regular speaker* S_{Reg} is almost identical to the S_1 speaker in standard RSA approaches. The main difference is that they take into account both social and semantic information in choosing the appropriate message. This has the effect of having them uttering a more socially marked message in case they should display a specific persona, *ceteris paribus*. If there is no incentive to display a specific persona, then they should choose the most informative message, à la RSA.

2. *The cooperative, diverse speaker* S_{Div} functions in the same way as their regular counterpart, but the key difference is that this player takes into account the diversity of the audience, and notably has access to multiple Literal Listeners, with differing priors. This leads to a difference in utility computation and thus different results and differences in message choice, that reflect the variety in the audience. This player is still considered *cooperative*, meaning that they have no intention of communicating different messages to the different audiences. It serves as a middle-ground between a regular speaker and a duplicitous speaker.

3. *The duplicitous speaker* S_{Dup} is the solution concept that is most likely to lead to the use dogwhistles in discourse. The specific difference between this speaker and the other two is the fact that the choice of message is no longer done simply based on a necessity to communicate one state of the world and one persona, but of communicating two states of the world and two personae, which are dependent upon the types of Literal Listener they are facing. Formally, the information they take as a basis for choosing messages,

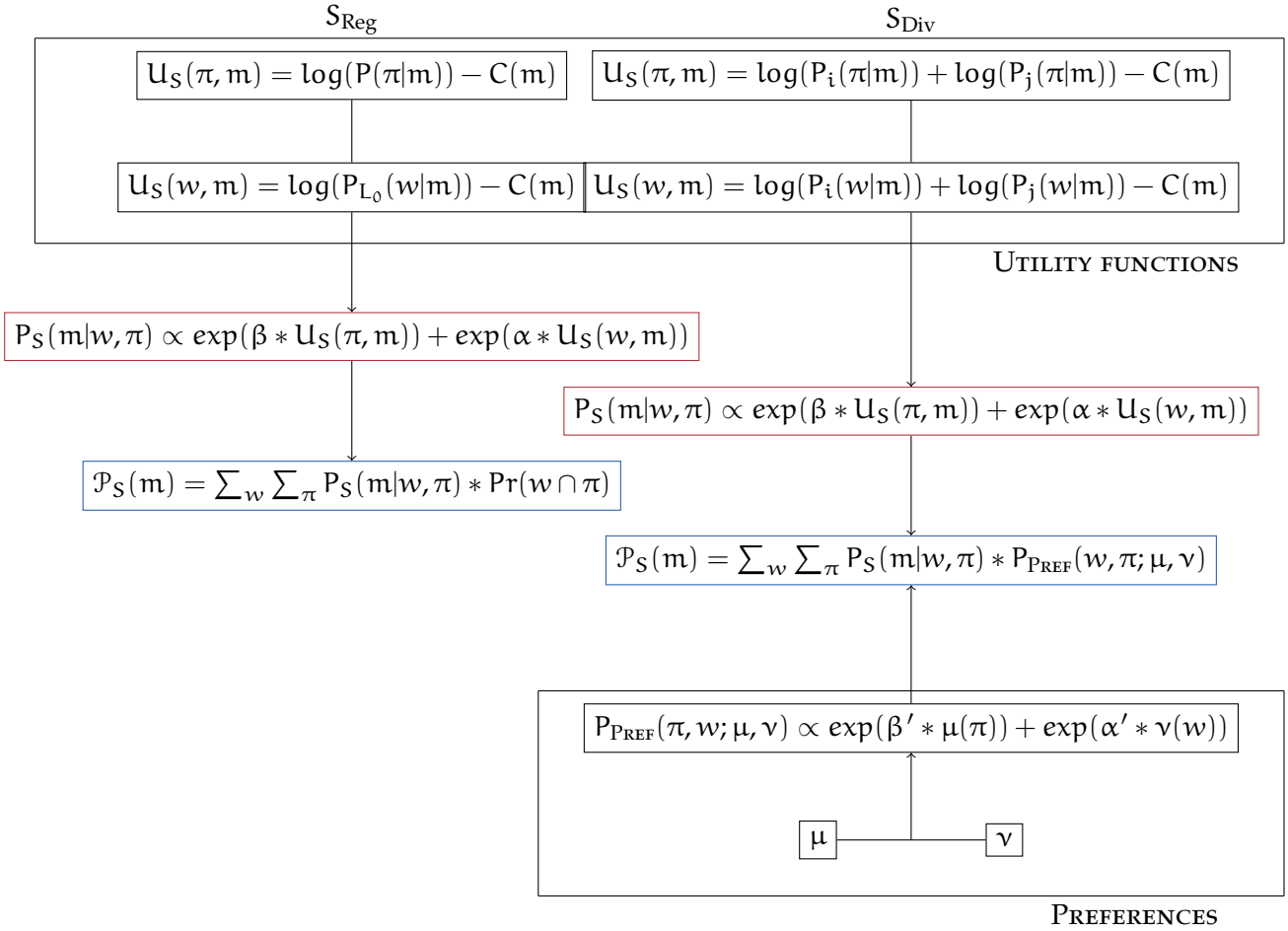


Figure 3: Overall view of the cooperative speaker solution concepts in DWG. The $\mathcal{P}(m)$ function gives the proportion of utterances that each message should represent/the probability that any given message will be chosen by the speaker. When no preferences are available, it should be based on a prior probability of any given $w \in W$ and $\pi \in \Pi$ to appear simultaneously. We have not defined this, but assuming independence between content and persona (an assumption likely to be wrong), we could compute it thus: $Pr(w \cap \pi) = Pr(w) * Pr(\pi)$. The red box is the *choice function*, which determines which utterance is more likely to be chosen given specific information to be transmitted. The function in the blue box allows us to have the probability, for each message, to be used.

instead of being of the form w, π, is of the form $\langle w_i, w_j \rangle, \langle \pi_i, \pi_j \rangle$, taking into account the maximization of information for Listeners L_i and L_j separately.

All those families can then be enriched using a *preference function* over personae, as per Burnett (2019), turning the game into a game with persona selection. This preference function in SMG settings does not actually change the utility of each option, but has an influence on the overall distribution of utterance selection. In addition to that preference function over personae, we have added a preference function over possible worlds to the *duplicitous speaker* S_{Dup}. The idea here is that being in a non-cooperative setting, there is no incentive for the speaker to convey the "true" state of the world (as per the maxim of quality). In short, the preference function over possible worlds openly allows for lying to happen.

All the relevant functions pertaining to the two cooperative speaker types are found in Figure 3 and the ones pertaining to the duplicitous speaker are in Figure 4.

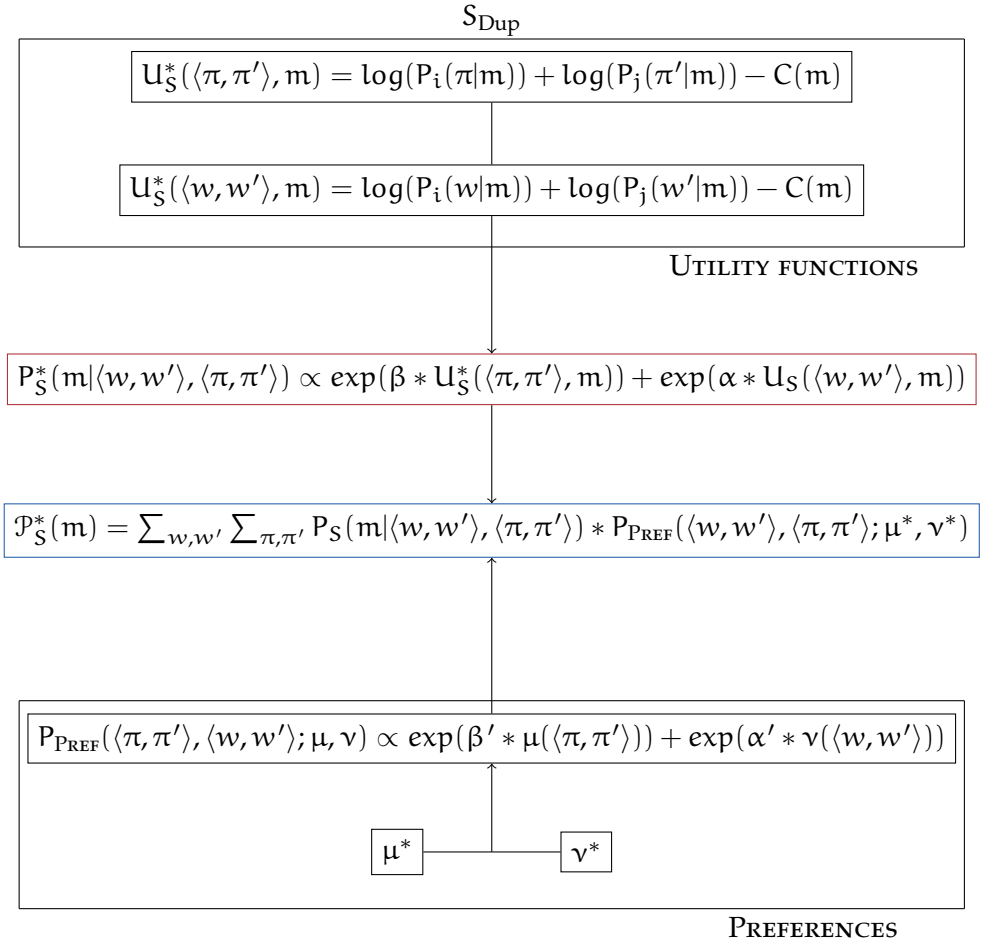


Figure 4: Overall view of the duplicitous speaker solution concept in DWG. The superscript stars in functions indicate that unlike cooperative speakers' functions, these do not take worlds or personae as arguments, but pairs of worlds and pairs of personae. The red box is the *choice function*, which determines which utterance is more likely to be chosen given specific information to be transmitted. The function in the blue box allows us to have the probability, for each message, to be used.

LISTENER FAMILIES:

1. *The pragmatic listener* L_1 is almost identical to its counterpart in standard RSA approaches. Like the Literal Listener in this model, it will have predictions over both personae and possible worlds after hearing an utterance. Pragmatic listeners envision the speaker as being a *cooperative, regular* speaker. This is key in our analysis of dogwhistles from a philosophical standpoint as being a case of abusing the mechanisms of the CP for non-cooperative goals.

2. *The cagey listener* L_{Cag} is a listener that envisions the speaker as being a duplicitous speaker. They come in two flavors, *simple cagey listener* and *uncovering cagey listener* (Burnett, 2019; Goffman, 1970). The *simple* version assumes a preference function for their envisioned speaker in order to interpret the utterance that they receive, whereas the *uncovering* version has priors over a set of possible preference functions that are updated upon hearing the utterance. Mirroring the *duplicitous speaker*, the cagey listener interprets pairs of possible worlds and personae instead of just one instance of each.

All the relevant functions pertaining to these listener solution concepts are found in Figure 5.

5.4 EXAMPLES

Taking into account only S_{Reg} and L_1, we have a game that is extremely similar to a more standard communication situation, involving the transmission of a single message between one speaker and one listener. We expect that in such a case, a dogwhistle term, being ambiguous, will generally not be preferred if there are less ambiguous counterparts available, much in the manner of how scalar implicatures work. Section 5.4.1 will explore how the model can, with minor tweaks, become a proper generalization of RSA and SMG models and give the exact same predictions in the same situations.

In order to properly model dogwhistle communication, we have to take into account that there is variation in the audience, not merely in their priors, but also in how they interpret language from both a social meaning and a semantics point of view. This was the reason behind the introduction of S_{Div}, a speaker that basically envisions two different Literal Listeners and uses both of their priors and supposed knowledge about language to choose their utterances. But we will see that this alone is insufficient to favor the use of a dogwhistle over more precise alternatives in Section 5.4.2.

Finally, in Section 5.4.3, we will use the example now discussed at length of "inner cities" and see what the model predicts once we inject S_{Dup} as a speaker. We will also see how L_{Cag} can be seen as an adequate representation of a listener that has the means to call out a dogwhistle upon hearing one.

L_1

$$L_1(w|m) \propto \sum_{\pi \in \Pi} P_S(m|w,\pi) * Pr(w)$$

$$L_1(\pi|m) \propto \sum_{w \in W} P_S(m|w,\pi) * Pr(\pi)$$

L_{Cag}

$$L_{Cag}(\langle w,w'\rangle|m) \propto \sum_{\langle \pi,\pi'\rangle} P^*_S(m|\langle w,w'\rangle,\langle \pi,\pi'\rangle) * P_{Pref}(\langle w,w'\rangle,\langle \pi,\pi'\rangle;\mu^*,\nu^*)$$

$$L_{Cag}(\langle \pi,\pi'\rangle|m) \propto \sum_{\langle w,w'\rangle} P^*_S(m|\langle w,w'\rangle,\langle \pi,\pi'\rangle) * P_{Pref}(\langle w,w'\rangle,\langle \pi,\pi'\rangle;\mu^*,\nu^*)$$

$\langle \pi,\pi'\rangle \in \Pi \times \Pi$

$\langle w,w'\rangle \in W \times W$

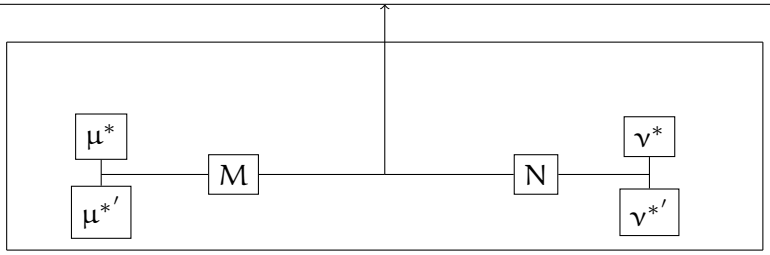
L^*_{Cag}

$$L^*_{Cag}(\langle w,w'\rangle,\nu^*|m) \propto \sum_{\mu^* \in M} \sum_{\langle \pi,\pi'\rangle} P^*_S(m|\langle w,w'\rangle,\langle \pi,\pi'\rangle) * P_{Pref}(\langle w,w'\rangle,\langle \pi,\pi'\rangle;\mu^*,\nu^*)$$

$$L^*_{Cag}(\langle \pi,\pi'\rangle,\mu^*|m) \propto \sum_{\nu^* \in N} \sum_{\langle w,w'\rangle} P^*_S(m|\langle w,w'\rangle,\langle \pi,\pi'\rangle) * P_{Pref}(\langle w,w'\rangle,\langle \pi,\pi'\rangle;\mu^*,\nu^*)$$

$\langle \pi,\pi'\rangle \in \Pi \times \Pi$

$\langle w,w'\rangle \in W \times W$



PREFERENCES PRIORS

Figure 5: Overall view of the Listener solution concepts in DWG. The output functions are in red boxes.

5.4.1 *Reproducing RSA and SMG*

Although our Literal Listener differs significantly from the standard in RSA, we will explore whether it can predict scalar implicatures the way standard RSA does. The model as is cannot do this, but with the addition of two specific parameters, the model becomes a proper generalization of both RSA and SMG.

RSA models derive successful communication (coordination on the right interpretation for each message in a set by the speaker and the listener) as deriving from informativity; speakers try to use the most informative statement possible, and listeners are aware of that, coordination ensues. If we look at the utility functions of the model presented in Figures 3 and 4, we see that these are exactly the same utility functions as what is found in RSA models, and in fact the whole speaker model is just like the conjunction of two different RSA models taking into account different kinds of informativity carried in messages. Supposedly, therefore, we should observe the same effects that we observe in other RSA models.

We start from the standard example of scalar implicatures derivation presented in Section 2.3.2. Scalar implicatures are standardly presented using quantifiers, words that carry little to no social meaning in themselves[9], we will do the same here and define our core parameters as follows, equivalent to what we had in Section 2.3.2:

Let Γ be a DWG: $\langle\{S_{Reg}, \{L_{1_i}, L_{1_j}\}\}, L_0, W, M, \Pi, \textsc{Lex}, \textsc{Soc}, \Pi\text{-}\textsc{Lex}, \Delta\text{-}\textsc{Soc}, Pr_W, Pr_\Pi\rangle$.

In the context of this example, the two listeners L_{1_i}, L_{1_j} are defined as identical (S_{Reg} can only take one L_0 into account anyway) and temperature parameters $\alpha, \alpha', \beta, \beta'$ are all set to 1[10].

- S_{Reg} is a regular speaker and L_{1_i}, L_{1_j} are identical pragmatic listeners.
- $L_0 : M \to \Delta(T), \Delta(\Pi)$ is a Literal Listener corresponding to the priors found below.
- $W = \{1, 2, 3\}$ is a set of possible states of the world.
- $M = \{\text{some, all}\}$ is the set of possible messages available to the speaker.
- $\Pi = \{\pi\}$ is the set of personae. In this context, we have to define one for the model to function, but we do not use it.
- $\textsc{Lex} = \{[\![\cdot]\!]\}, \textsc{Soc} = \{[\cdot]\}$ again, because of the simple situation here, there is just one interpretation function $[\![\cdot]\!]$ and one indexation function $[\cdot]$.
- Because of the configuration we have, $\Pi\text{-}\textsc{Lex}(\pi)$ will give $[\![\cdot]\!]$ with probability 1, and $\Delta\text{-}\textsc{Soc}([\cdot]) = 1$.
- The priors over worlds Pr_W are uniform. The priors over personae Pr_Π are uniform as well ($Pr_\Pi(\pi) = 1$).

9 Although their pronunciation might, of course.
10 Which is equivalent to not having them at all.

$[\![\cdot]\!]$	1	2	3
$[\![\text{some}]\!]$	1	1	1
$[\![\text{all}]\!]$	0	0	1

Table 12: The interpretation function for scalar implicatures game analyzed through the lens of the DWG framework. This is identical to Table 3. It is also displayed here in Boolean form.

$P_{L_0}(w\|m)$	1	2	3
some	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
all	0	0	1
$P(\pi\|m)$	π		
some	1		
all	1		

Table 13: The literal listener for the scalar implicatures game analyzed through the lens of the DWG framework.

$[\![\cdot]\!]$ is defined in Table 12. Let's first look at what the model with no modification does. Table 13 gives us the relevant interpretations for the literal listener. We can see that they are the same as what we have in standard RSA. Importantly, however, there is another interpretation function at the Literal Listener level in our model: $P(\pi|m)$. This gives an odd but clearly predictable result: whatever the message sent, the social meaning is the same. Since we defined our Π set as containing only one π, this is understandable. Whatever the message, there is only one available persona anyway, so anyone saying anything will display that persona.

In addition to being uninteresting as a result, this feature of the model proves to be catastrophic if we try to derive simple scalar implicatures. Indeed, applying the computations presented in Figure 3, we reach the results found in Table 14. We can see that the model makes absurd predictions, whereby S_{Reg} is somewhat likely to use "all" in contexts where it is obviously untrue. Although we do have the right tendency ("some" is less likely to be used to mean "all" than any other thing), these predictions are unsettling. What happened here? Well, speakers in the DWG model do not choose to utter their messages based only on *content*, they also, systematically, take into account social meaning. There's the hiccup: all messages are compatible with persona π. Because the speaker necessarily has a persona to convey, this leads to no words being incompatible with what it is they want to say, no matter the state of the world. The pragmatic listener L_1 takes the cooperative speaker as a basis for their reasoning and so falls into the same trap. As we multiply the layers of recursive thinking, this gets worse and players slowly descend into madness as their language devolves into unsignifying sound and fury, as illustrated by Figure 6.

This is obviously terrible, but the good news is we can do something about it: introducing *sensitivity parameters* $\sigma, \tau \in \mathbb{R}$. These two numerical parameters allow us to quickly transform the DWG model into a proper generalization of both RSA and SMG. The only thing we need

$P_S(m|w,\pi)$	SOME	ALL
$1,\pi$	≈ 0.571	≈ 0.429
$2,\pi$	≈ 0.571	≈ 0.429
$3,\pi$	0.4	0.6

Table 14: Results for S_{Reg} in the scalar implicatures derivation in DWG.

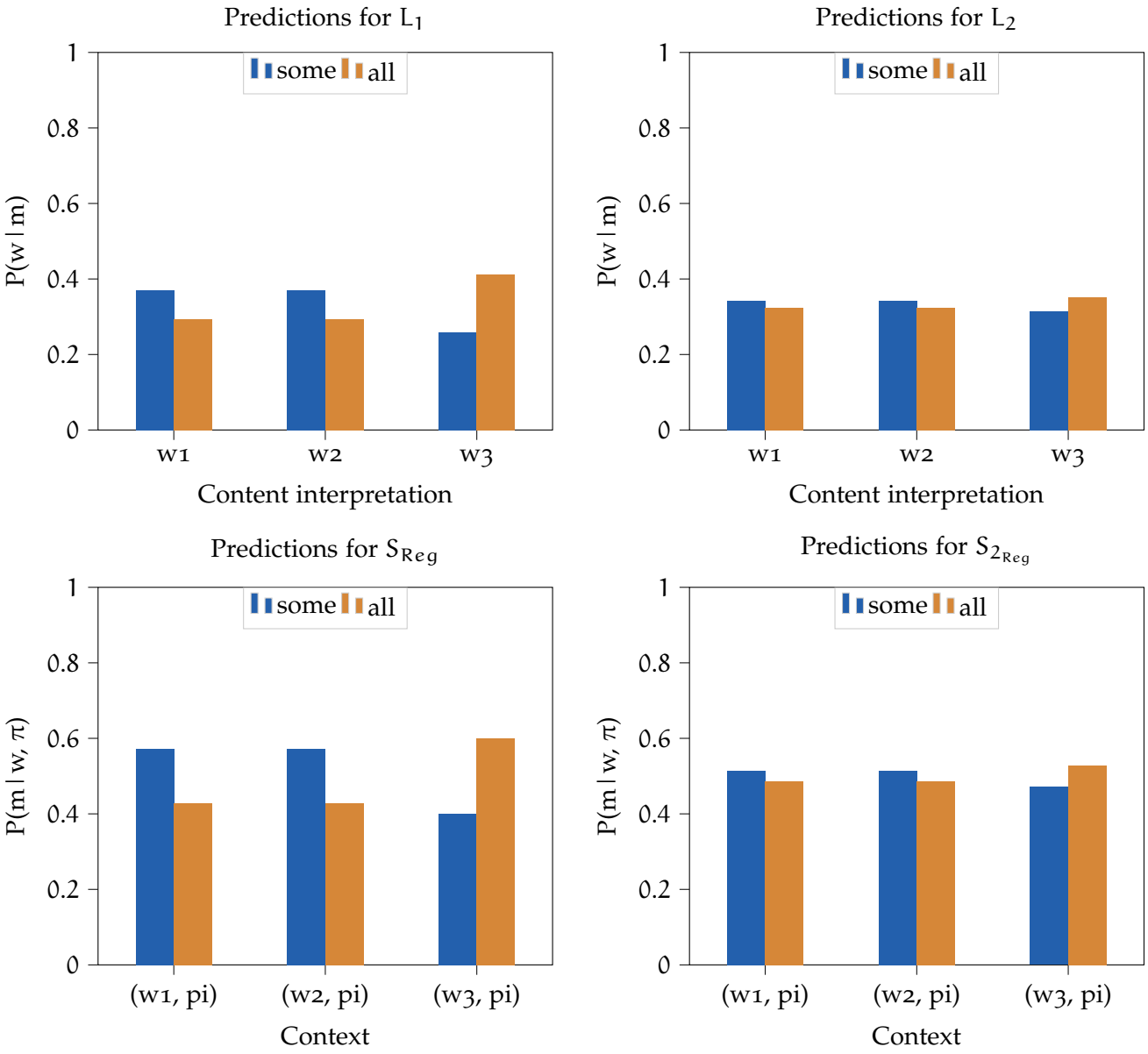


Figure 6: Visualization of speakers and listeners for the derivation of scalar implicatures in DWG. on the left are S_{Reg} and L_1, and on the right would be the equivalent of S_2 and L_2, recursively defined from our existing solution concepts.

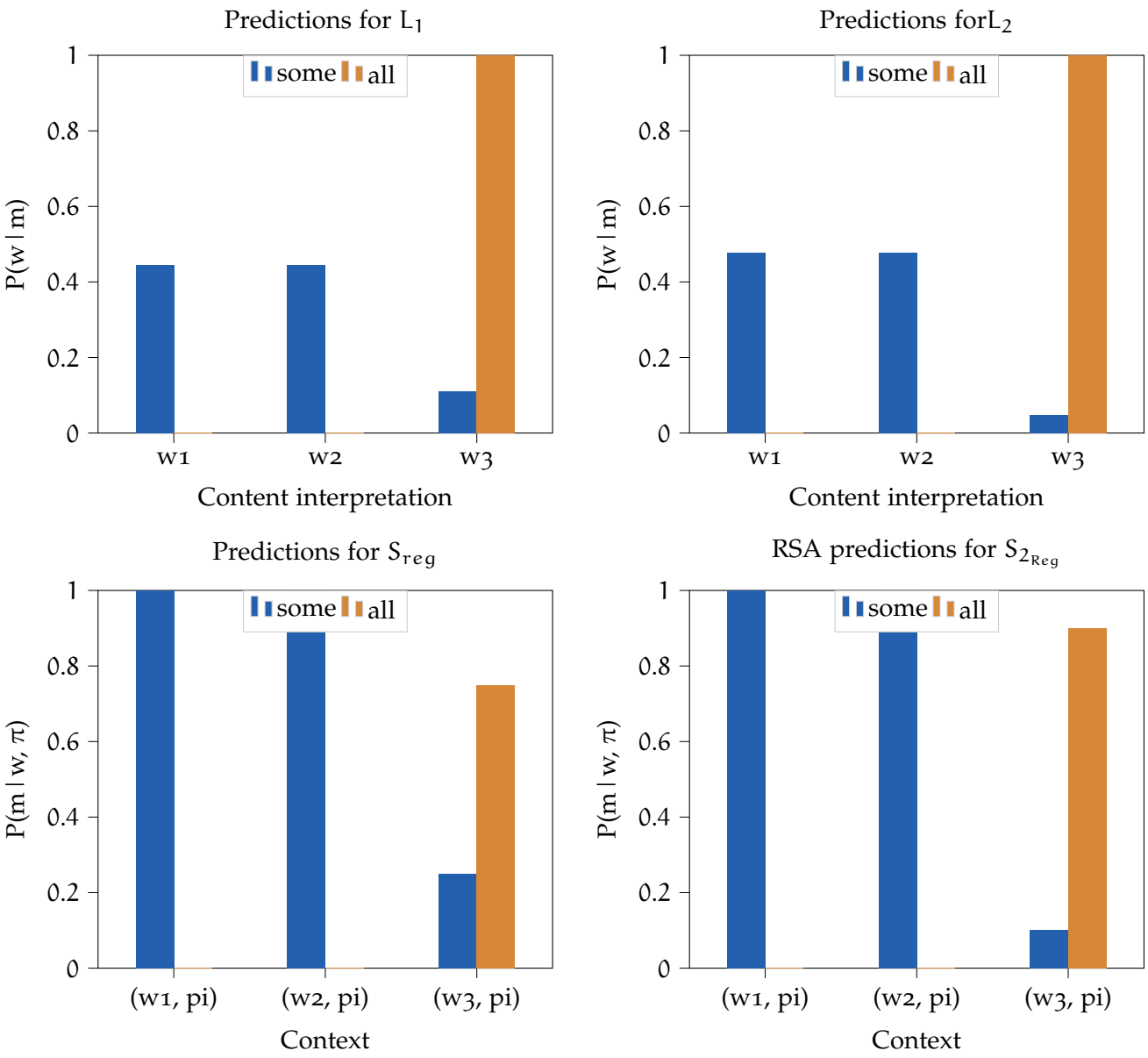


Figure 7: Visualization of speakers and listeners for the derivation of scalar implicatures in DWG. on the left are S_{Reg} and L_1, and on the right would be the equivalent of S_2 and L_2, recursively defined from our existing solution concepts, with $\sigma = 0, \tau = 1$. Order has been restored and all is well.

to do is modify the choice rules of our speakers in the following way (taking S_{Reg} as an example):

$$P_S(m|w,\pi) \propto exp(\beta * U_S(\pi, m)) + exp(\alpha * U_S(w, m))$$
$$\mapsto P_S(m|w,\pi) \propto \sigma * exp(\beta * U_S(\pi, m)) + \tau * exp(\alpha * U_S(w, m))$$

If we now set $\sigma = 0$ and $\tau = 1$, we have completely nullified the influence of the social meaning component of our messages. See Figure 7 to see order being gradually restored.

This is obvious once it is implemented. Because DWG is a mere conjunction of RSA and SMG at that step, it goes without saying that if we suppress the SMG part of the model, then we end

up with RSA. Symmetrically, setting $\sigma = 1, \tau = 0$, and defining our variables and indexation functions the way they are defined in the Obama example in Burnett (2019), we get the exact same results as the paper. So adding the sensitivity parameters σ and τ allows us to turn DWG into a proper generalization of both RSA and SMG. Using different configurations for both parameters, we can even simulate and endless amount of situations where social and content meaning play different roles (keeping in mind only the ration between the two counts, not the actual values). While satisfying, this tweak feels inappropriate.

Indeed, one key insight behind DWG is that there is no way to communicate content without also communicating social meaning. Nevertheless, in order to reproduce the results from both RSA and SMG, we need such a suppression mechanism. The choice of which σ, τ we have can be based on the social situation[11] or individual-based. How to determine those is very likely to be outside the scope of linguistics, and in any case is definitely outside the scope of this dissertation, although Henderson and McCready (2018) presents a similar feature which is attached to specific personalities.

5.4.2 *Adding* S_{Div}

So far we have only considered cases that are not characterized by the term *dogwhistles*. Typically, and as discussed in the previous chapter, dogwhistles are likely to arise in situations of political conflict where the speaker aims to satisfy a set of preferences and not merely convey information. Let us start already by considering heterogeneous crowds as listeners. In DWG this is simply made evident by having an additional listener in the set of players. To represent the political conflict/ideological heterogeneity we give that listener different priors. The speaker, in turn, takes into account this heterogeneity in the audience when deciding what message to utter. The first thing we are going to do is implement such as speaker, though we will first remain in the realm of cooperation, to see whether it is predicted that dogwhistles could arise without manipulative intent. The speaker we are envisioning here, S_{Div}, differs from a proper dogwhistle utterer like S_{Dup} in the sense that it is still considered to be cooperative to some extent. This translates into this speaker having no preferences over worlds or personae, but also only wishing to communicate a single message to the audience. It is to be read as an agent merely communicating the state of the world as it is, like a standard RSA speaker would. Arguably, this means very little in the context of political speech, for example, but it is nonetheless a starting point.

We will start here to develop an example that we will re-use until the end of the chapter, inspired by the now familiar "inner cities" dogwhistle[12], specifically in the version uttered by Paul Ryan: (15a). To do so we will work with the following:

11 In some cases it might be the case that people do indeed care less about social signaling than in other cases, a typical theoretical example would be, e. g., scientific discourse, but I do not think this statement would hold to careful observation. My guess is that such situations are purely hypothetical.

12 Although viewed from a much simpler point of view that in many ways would not be in keeping with what we presented from Saul (2018b) and Khoo (2017), for example.

$[m]$	$[\cdot]_{rc}$	$[\cdot]_{nrc}$
m_{cc}	$\{\pi_{nrc}\}$	$\{\pi_{nrc}\}$
m_{aa}	$\{\pi_{rc}\}$	$\{\pi_{rc}\}$
m_{dw}	$\{\pi_{rc}\}$	$\{\pi_{nrc}\}$

Table 15: Definition of the indexation functions $[\cdot] \in$ Soc. We assume that an overtly racist content leads to a racist persona, addressing the slight philosophical concern raised in Section 5.3.1.1.

Let Γ be a DWG:

$$\langle \{S_{Div}, \{L_i, L_o\}\}, L_{0_{i/o}}, W, M, \Pi, \textsc{Lex}, \textsc{Soc}, \Pi\text{-}\textsc{Lex}_{i/o}, \Delta\text{-}\textsc{Soc}_{i/o}, Pr_{W_{i/o}}, Pr_{\Pi_{i/o}} \rangle,$$

where:

- $M = \{$"city centers", "African-American neighborhoods", "inner cities"$\}$ our set of possible messages contains three messages. In tables, for readability, these will be referred to as m_{cc}, m_{aa}, m_{dw}. One of those, "inner cities", is our dogwhistle term. The other two are supposedly unambiguous versions of each of the possible meanings of "inner cities": "city centers" is compositionally equivalent and should have the same reference as "inner cities" in a dogwhistle-less world; "African-American neighborhoods" is the explicit version of what is supposed to be the dogwhistled content of "inner cities".

- $W = \{w_{aa}, w_{cc}\}$ there are two available interpretations: one where we are referring to African-American neighborhoods (w_{aa}) and one where we are not (w_{cc}).

- $\Pi = \{\pi_{rc}, \pi_{nrc}\}$ there are two relevant personae that are considered: "racist conservative" (π_{rc}) and "non-racist conservative" (π_{nrc}). These personae stand for ideological groups. In that respect they are rather different from what is found in Burnett (2019), where personae are based on individual properties and character traits. Here, we take the liberty of adopting a more First Wave approach to the question and have variation refer to an entire group. The properties that we refer to by using these names are properties of social groups and ideologies more than proper individuals, but in the case of political communication it does not seem to us out of place to do so.

- $\textsc{Soc} = \{[\cdot]_{rc}, [\cdot]_{nrc}\}$ there are two indexation functions, corresponding to the two personae we have defined.

- $\textsc{Lex} = \{[\![\cdot]\!]_{rc}, [\![\cdot]\!]_{nrc}\}$ there are similarly two interpretation functions, one for each persona.

All $[\cdot] \in \textsc{Soc}$, $[\![\cdot]\!] \in \textsc{Lex}$ are defined in Tables 15 and 16.

We then define two listeners, L_i and L_o, respectively thought of as the ingroup and outgroup listener, with the following priors:

$[\![m]\!]$	$[\![\cdot]\!]_{rc}$	$[\![\cdot]\!]_{nrc}$
m_{cc}	$\{w_{cc}\}$	$\{w_{cc}\}$
m_{aa}	$\{w_{aa}\}$	$\{w_{aa}\}$
m_{dw}	$\{w_{aa}, w_{cc}\}$	$\{w_{cc}\}$

Table 16: Definition of the interpretation functions $[\![\cdot]\!] \in$ LEX.

- For each, a uniform prior distribution over worlds (meaning neither have a priori knowledge about the content of what is about to be said to them).
- For each, a uniform prior distribution over personae (meaning neither have a priori knowledge about the identity of the speaker[13]).
- We assume in that case that each speaker only has access to one indexation function $[\cdot]$, that of their preferred group. This translates into the Δ-SOC function for, say, L_i to necessarily interpret messages using the $[\cdot]_{rc}$ indexation function.

  Because we only have two personae available, we have:

$$\Delta\text{-SOC}(\pi_{rc}) = p$$
$$\Delta\text{-SOC}(\pi_{nrc}) = 1 - p$$

  With $p = 1$ for L_i and $p = 0$ for L_o.

- Regarding Π-LEX, things are a little bit different. To underline the discrepancy between members of the public, and specifically the ignorance of outgroup members, we have the following:

  For L_i:

$$\Pi\text{-LEX}(\pi_{rc}) = \begin{cases} [\![\cdot]\!]_{rc} : 1 \\ [\![\cdot]\!]_{nrc} : 0 \end{cases}$$
$$\Pi\text{-LEX}(\pi_{nrc}) = \begin{cases} [\![\cdot]\!]_{rc} : 0 \\ [\![\cdot]\!]_{nrc} : 1 \end{cases}$$

  For L_o:

13 This is obviously highly unlikely to be the case, especially in the context of politics, but one point that it illustrates is that, as we will see, dogwhistle interpretations can arise even if the speaker is completely unknown from the crowd, thus acting as the proverbial "secret handshake" we've already mentioned multiple times.

$P_{L_0}(w\|m)$	w_{aa}	w_{cc}
m_{cc}	0	1
m_{aa}	1	0
m_{dw}	0.5	0.5
$P(\pi\|m)$	π_{rc}	π_{nrc}
m_{cc}	0	1
m_{aa}	1	0
m_{dw}	1	0

Table 17: Literal listener results for listener L_i.

$P_{L_0}(w\|m)$	w_{aa}	w_{cc}
m_{cc}	0	1
m_{aa}	1	0
m_{dw}	0	1
$P(\pi\|m)$	π_{rc}	π_{nrc}
m_{cc}	0	1
m_{aa}	1	0
m_{dw}	0	1

Table 18: Literal listener results for listener L_o.

$$\Pi\text{-Lex}(\pi_{rc}) = \begin{cases} [\![\cdot]\!]_{rc} : 0 \\ [\![\cdot]\!]_{nrc} : 1 \end{cases}$$

$$\Pi\text{-Lex}(\pi_{nrc}) = \begin{cases} [\![\cdot]\!]_{rc} : 0 \\ [\![\cdot]\!]_{nrc} : 1 \end{cases}$$

Now that we have all this set, we can see what the predictions for Literal Listeners are, these are presented in Tables 17 and 18. They behave exactly as our interpretation of dog-whistles predicts, meaning that the dogwhistle term is treated as identical to the acceptable unambiguous version of it by L_o but treated differently by L_i, who clearly interprets it as an unambiguous social signal while at the same time being unsure of its interpretation in terms of content.

We now try to use S_{Div} as a speaker, it uses both existing Literal Listeners to compute its preferences over utterances. The results of $P_S(m|w,\pi)$ when using S_{Div} as a speaker are found in Table 19. Again, it seems that our system adds in uncertainty, for the same reasons as before. We do however observe the right tendencies: in order to communicate $\langle w_{aa}, \pi_{rc} \rangle$, the

$P_S(m\|w,\pi)$	m_{cc}	m_{aa}	m_{dw}
w_{aa},π_{rc}	0	1	0
w_{aa},π_{nrc}	0.5	0.5	0
w_{cc},π_{rc}	0.4	0.4	0.2
w_{cc},π_{nrc}	0.8	0	0.2

Table 19: Results for S_{Div}.

$P_{L_1}(w\|m)$	w_{aa}	w_{cc}
m_{cc}	≈ 0.269	≈ 0.731
m_{aa}	$0.7\overline{72}$	$0.2\overline{27}$
m_{dw}	0.5	0.5
$P(\pi\|m)$	π_{rc}	π_{nrc}
m_{cc}	≈ 0.192	≈ 0.808
m_{aa}	$0.68\overline{18}$	$0.3\overline{18}$
m_{dw}	$0.68\overline{18}$	$0.3\overline{18}$

Table 20: Pragmatic listener results for listener L_i.

only message that the speaker might choose is "African-American neighborhoods", a message that is judged to be strictly incompatible with the communication of $\langle w_{nr}, \pi_{nrc}\rangle$. The strangest situation is the communication of $\langle w_{aa}, \pi_{nrc}\rangle$, but this is expected, we have set the content and social meanings of our messages such that carrying a content that clearly indicates racist conservative views will invariably index a racist conservative persona, it is therefore just about impossible for that speaker to communicate specifically this. Note too that in that situation, the dogwhistle isn't even considered.

Because of its vagueness, the dogwhistle is never the preferred option for that speaker. This comes from the fact that it is still envisioned as being, in a way, a cooperative speaker. S_{Div} is therefore not a good speaker solution concept if we are interested in the production of dogwhistles, since it is predicted to almost never use them, including in seemingly appropriate situations. If we look at Pragmatic Listeners in this model, for both L_i and L_o, we have interesting results, which can be seen in Tables 20 and 21. We again observe that there is some uncertainty compared with the Literal Listeners. This is however not necessarily an issue in a dogwhistle situation, since if we recall from the properties that we want our model to have, this is compatible with PLAUSIBLE DENIABILITY.

An issue that arises however, is that as before, if we repeat the recursive thinking over several steps, it will once more be the case that we end up with meaningless messages. Again, this is fixable using σ and τ set at the appropriate levels. In this case, however, the fact that we want plausible deniability also means that we do not want our predictions to converge on strict interpretations. We also have to add here that because they have access to both the lexicon of their own group and of the outgroup, the Pragmatic Listener for L_i sees their uncertainty

$P_{L_1}(w\|m)$	w_{aa}	w_{cc}
m_{cc}	≈ 0.286	≈ 0.714
m_{aa}	0.8	0.2
m_{dw}	≈ 0.286	≈ 0.714
$P(\pi\|m)$	π_{rc}	π_{nrc}
m_{cc}	≈ 0.286	≈ 0.714
m_{aa}	0.8	0.2
m_{dw}	≈ 0.286	≈ 0.714

Table 21: Pragmatic listener results for listener L_o.

increase even more at each step. This can be seen in Figures 8 and 9. Note that the Pragmatic Listener here has not been changed since the last example, meaning that their image of the speaker is the cooperative non-diverse speaker based on their priors and *not* S_{Div}.

So what about the Pragmatic Listener here? Are its results undesirable? Yes and no, what we have here is a set of Literal Listeners that behave in a way that is *exactly* what dogwhistles predict, so in a way, the Literal Listener here might be a good model in and of itself of theoretical listeners in a dogwhistle context. The Pragmatic Listeners can in themselves be thought of as higher level listeners, for whom the interpretation of messages is no longer clear due to the known variability in the choices of the speaker. In any case however, the most probable interpretations that the Pragmatic Listeners make of each possible message are in keeping with what we are expecting from the model, and their apparent oddness is in fact in keeping with the idea that speakers communicate more than just semantics when they use language. Importantly, the Pragmatic Listeners that we have here allow us in part to explain the *unintended dogwhistles* effect described by Saul (2018a): there is no need for the speaker to communicate the dogwhistled content intentionally when using a dogwhistle (say, by repeating it) for it to be interpreted by listeners who have the right priors.

5.4.3 S_{Dup}

If we look at Pragmatic Listeners' interpretations of the dogwhistle term in Tables 20 and 21, we do see a discrepancy, and the dogwhistle is more likely to be interpreted as meaning w_{aa} by L_i and more likely to be understood as meaning w_{cc} by L_o. Regarding personae, we also observe that L_i associates π_{rc} to the speaker when L_o does not. So on the Listener part, using our version of a Literal Listener, a speaker S_{Reg} and the priors we've defined, we already have a dogwhistle effect. What we need is a speaker that will use dogwhistles when it is appropriate, when both S_{Reg} and S_{Div} had very few chances of ever using a dogwhistle no matter the situation.

S_{Div} was not a rich enough speaker model to accurately describe dogwhistle communication. One important point about it is that even though this speaker takes into account the

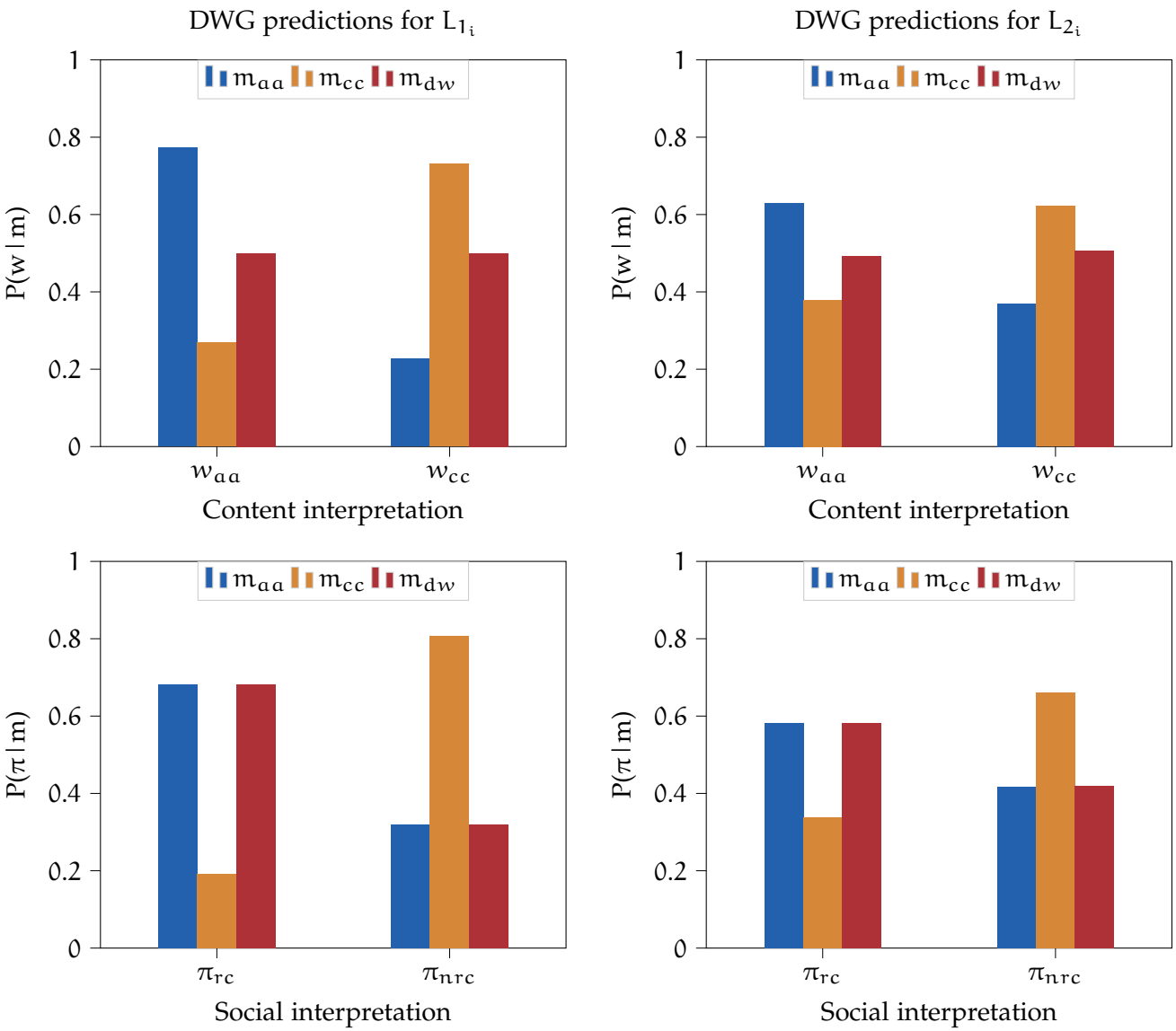


Figure 8: Visualization of listeners L_i in our DWG example. Uncertainty increases with added layers of recursive meaning. Because they have more knowledge about the other group's lexicon, L_i listeners see their uncertainty increasing faster. L_1 are on the left, the equivalent to recursive L_2 on the right.

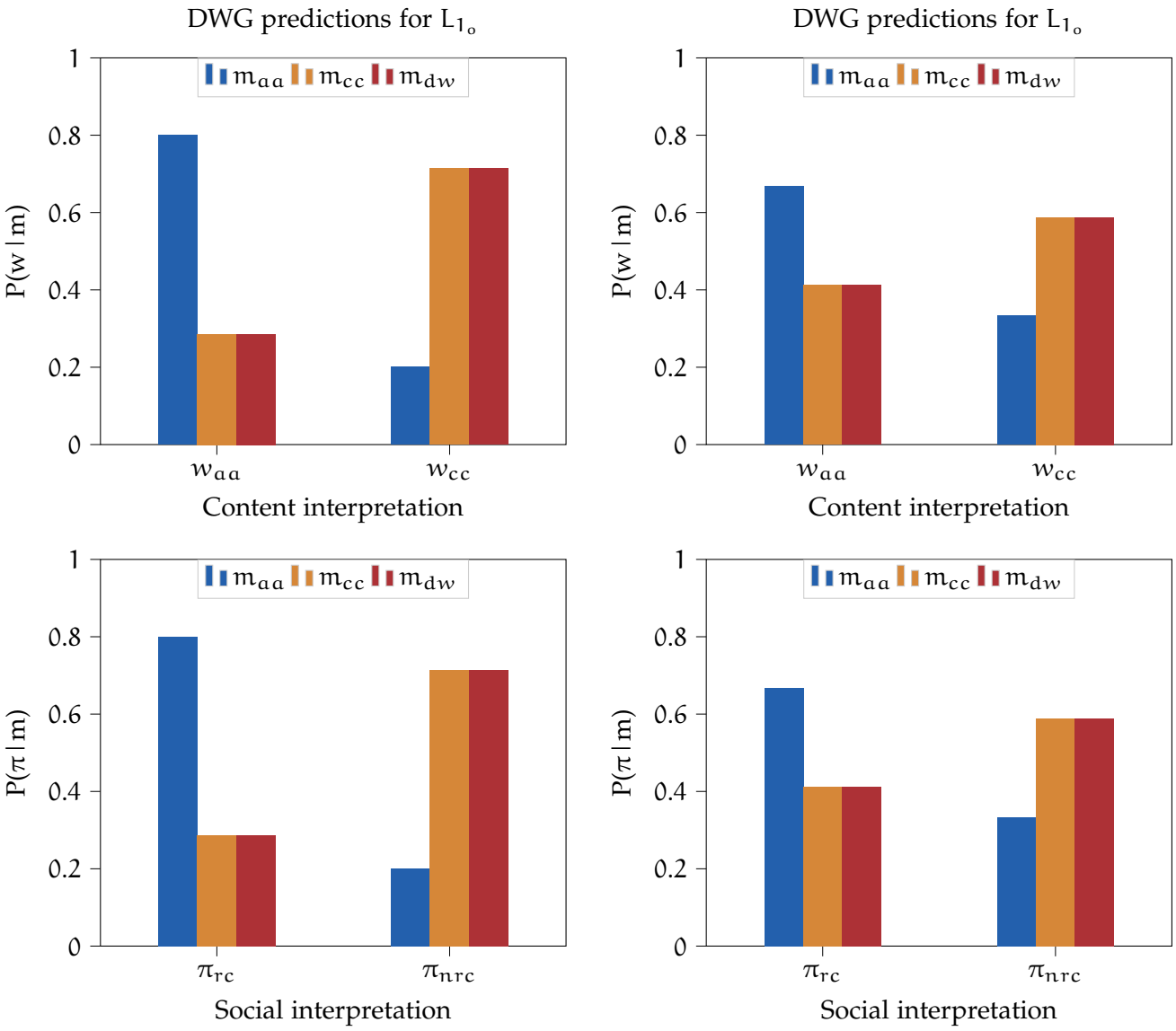


Figure 9: Visualization of listeners L_o in our DWG example. Uncertainty increases with added layers of recursive meaning. L_1 are on the left, the equivalent to recursive L_2 on the right.

variety of priors in the audience, they do not focus at all on the idea of *coded message*, they do not specifically communicate a different message to each group in the audience. S_{Dup} should address this issue.

In the case of S_{Dup}, the key change is that beyond having access to two different Literal Listeners, they have also the possibility of choosing a message by treating them as independent entities, assuming that they are communicating two messages since there are two listeners. This translates into every function of the solution concept no longer using worlds and personae w, π, but pairs of worlds and personae $\langle w, w' \rangle, \langle \pi, \pi' \rangle$. Among other things, this leads to them considering many more possible situations. Where S_{Reg} or S_{Div} would condition on 4 possible cases in choosing their message ($|W \times \Pi|$), S_{Dup} conditions over 16 cases ($|W \times W \times \Pi \times \Pi|$). Among those cases there will be some that are cases of coded language, for example any case involving $\langle w_{aa}, w_{cc} \rangle$ or $\langle \pi_{rc}, \pi_{nrc} \rangle$, where the message to be communicated should trigger different interpretations according to the listener.

The predictions for S_{Dup} are found in Table 22. We can see two things: first of all, S_{Dup} replicates S_{Div} in non-duplicitous situations, that is situations where the message to be communicated to both listeners are the same, making it a generalization of S_{Div}. We also see that S_{Dup} successfully chooses the dogwhistles in situations where the dogwhistle term is intuitively the best option to communicate a coded message. There is however one problematic point in this: the context defined by the speaker meaning to say $\langle w_{cc}, w_{aa} \rangle, \langle \pi_{nrc}, \pi_{rc} \rangle$ is undefined in this model. Due to the way utilities are computed, we end up with a division by zero when applying the $S_{Dup}(m|\langle w, w' \rangle, \langle \pi, \pi' \rangle)$ choice rule using this context. If we look more closely at this example, it constitutes the exact inverse of the dogwhistle situation that we have defined. What this tells us is that our set of messages does not contain any element that would be valid in that situation. Given our messages, their social meanings, their semantics, this situation is outside the scope of the model. Given that there are no other possible messages in our model, this is translated into an error, a mathematical absurdity. To get an intuition for it, it would be like trying to discuss the color of an object while only having access to words referring to shapes.

5.4.3.1 *Preferences*

As per Burnett (2019), we can have a measure of how likely it is that S_{Dup} uses the dogwhistle at all in any kind of interaction with those specific listeners by computing $\mathcal{P}(m)$[14]:

$$\mathcal{P}(m_{cc}) \approx 0.3714$$

$$\mathcal{P}(m_{aa}) \approx 0.3143$$

$$\mathcal{P}(m_{dw}) \approx 0.3143$$

14 We do so by treating the anomalous case as giving a probability of 0 for all messages.

$P_S^*(m\|\langle w, w'\rangle\langle \pi, \pi'\rangle)$	m_{aa}	m_{cc}	m_{dw}
$\langle w_{aa}, w_{aa}\rangle\langle \pi_{rc}, \pi_{rc}\rangle$	1	0	0
$\langle w_{aa}, w_{aa}\rangle\langle \pi_{rc}, \pi_{nrc}\rangle$	0.5	0	0.5
$\langle w_{aa}, w_{aa}\rangle\langle \pi_{nrc}, \pi_{rc}\rangle$	1	0	0
$\langle w_{aa}, w_{aa}\rangle\langle \pi_{nrc}, \pi_{nrc}\rangle$	0.5	0.5	0
$\langle w_{aa}, w_{cc}\rangle\langle \pi_{rc}, \pi_{rc}\rangle$	$0.\overline{66}$	0	$0.\overline{33}$
$\langle w_{aa}, w_{cc}\rangle\langle \pi_{rc}, \pi_{nrc}\rangle$	0	0	1
$\langle w_{aa}, w_{cc}\rangle\langle \pi_{nrc}, \pi_{rc}\rangle$	0	0	1
$\langle w_{aa}, w_{cc}\rangle\langle \pi_{nrc}, \pi_{nrc}\rangle$	0	$0.\overline{66}$	$0.\overline{33}$
$\langle w_{cc}, w_{aa}\rangle\langle \pi_{rc}, \pi_{rc}\rangle$	1	0	0
$\langle w_{cc}, w_{aa}\rangle\langle \pi_{rc}, \pi_{nrc}\rangle$	0	0	1
$\langle w_{cc}, w_{aa}\rangle\langle \pi_{nrc}, \pi_{rc}\rangle$	?	?	?
$\langle w_{cc}, w_{aa}\rangle\langle \pi_{nrc}, \pi_{nrc}\rangle$	0	1	0
$\langle w_{cc}, w_{cc}\rangle\langle \pi_{rc}, \pi_{rc}\rangle$	0.4	0.4	0.2
$\langle w_{cc}, w_{cc}\rangle\langle \pi_{rc}, \pi_{nrc}\rangle$	0	0.4	0.6
$\langle w_{cc}, w_{cc}\rangle\langle \pi_{nrc}, \pi_{rc}\rangle$	0	$0.\overline{66}$	$0.\overline{33}$
$\langle w_{cc}, w_{cc}\rangle\langle \pi_{nrc}, \pi_{nrc}\rangle$	0	0.8	0.2

Table 22: Results for S_{Dup}. The preferred message for each situation is in blue, the situation that is closest to our description of dogwhistles is in red. The red box indicates the anomalous case discussed in Section 5.4.3.

Without preferences (or, alternatively, setting the preference functions such that all possible contexts have the same value), we can see that the mere fact that there are listeners with different beliefs is not enough for the dogwhistle to be the term most likely to be used. This is because RSA-based models favor messages that are more informative, and there are more cases where, e. g. m_{cc} is more informative than m_{dw}. One way to solve this is to add to S_{Dup} preferences over personae (as in Burnett, 2019) and over worlds (which is licensed here by the fact that we are in a non-cooperative context not bounded by the communication of truth).

We define two preference functions $\mu^* : \Pi \times \Pi \mapsto \mathbb{R}$, $\nu^* : W \times W \mapsto \mathbb{R}$ that characterize a speaker's possible preferences over what they want to communicate. A possible set of μ^*, ν^* compatible with a speaker that wishes to communicate a coded message like we've been talking about is presented in Table 23. Using these preference functions, we now obtain:

$$\mathcal{P}(m_{cc}) \approx 0.2547$$
$$\mathcal{P}(m_{aa}) \approx 0.1975$$
$$\mathcal{P}(m_{dw}) \approx 0.5271$$

Given these preferences, the dogwhistle message is now much more likely to be used than any of the other alternatives, making S_{Dup} with preferences an appropriate speaker model in dogwhistle interactions.

$\langle w, w' \rangle$	μ^*
$\langle w_{aa}, w_{aa} \rangle$	0
$\langle w_{aa}, w_{cc} \rangle$	2
$\langle w_{cc}, w_{aa} \rangle$	0
$\langle w_{cc}, w_{cc} \rangle$	1
$\langle \pi, \pi' \rangle$	ν^*
$\langle \pi_{rc}, \pi_{rc} \rangle$	0
$\langle \pi_{rc}, \pi_{nrc} \rangle$	2
$\langle \pi_{nrc}, \pi_{rc} \rangle$	0
$\langle \pi_{nrc}, \pi_{nrc} \rangle$	1

Table 23: Preference functions for S^*_{Dup}. Any case that leads to the outgroup speakers being aware of the hidden meaning/social identity is dispreferred, cases leading to the ingroup understanding the hidden message while the outgroup remaining unaware of it are preferred.

Our speaker is now more likely to use dogwhistles than other messages, making S_{Dup} with preferences an appropriate speaker model in dogwhistle interactions. This concept of preference function is presented in Burnett (2019) as a way to take SMG to the Third Wave of variationist studies by acknowledging that personae are not merely transmitted by speakers based on any actual persona that they might have and wish to convey, but are instead actively constructed in conversations. This notion makes very little sense in signaling games as we have presented them in, e. g., Appendix A. In general, the notion makes very little sense in cooperative communication games, where the content of the message is pre-determined/dictated by the state of the world. In situations like dogwhistles, however, and one could argue in situations of non-cooperative communication in general, the incentive to respect Grice's maxim of quality is not as strong, or might even be non-existent. In those cases, as is the case in a Third Wave approach to identity construction, there is no "real state of the world" to communicate, there are a number of messages that have had their signification constructed over repeated interactions over time that are then used with their conventional meaning to convey whatever the speaker wants to convey.

In the case of dogwhistles, what we did is that we took the "coded message" interpretation in a rather literal way, with a definite intention from speakers to conceal something from the general audience. This is why we have envisioned a listener that uses the same kind of cooperative solution concepts that we had already presented. S_{Dup}, however, is not using the same kind of solution concepts at all, and as is the case for SMG, a notion of preference for identities and content may lead to a better characterization of the behavior of a speaker unconstrained by a supposed state of nature. We have effectively made an uncooperative speaker that interacts with a set of cooperative listeners.

But what about the possibility of an uncooperative listener? What would their preferred actions be in such a game?

$L_{Cag}(\langle w, w' \rangle \mid m)$	m_{aa}	m_{cc}	m_{dw}
$\langle w_{aa}, w_{aa} \rangle$	≈ 0.5921	≈ 0.1128	$0.\overline{09}$
$\langle w_{aa}, w_{cc} \rangle$	≈ 0.1316	≈ 0.1504	$0.\overline{48}$
$\langle w_{cc}, w_{aa} \rangle$	≈ 0.1974	≈ 0.2256	$0.\overline{18}$
$\langle w_{cc}, w_{cc} \rangle$	≈ 0.0789	≈ 0.5113	$0.\overline{24}$
$L_{Cag}(\langle \pi, \pi' \rangle \mid m)$			
$\langle \pi_{rc}, \pi_{rc} \rangle$	≈ 0.6052	≈ 0.0902	$0.0\overline{96}$
$\langle \pi_{rc}, \pi_{nrc} \rangle$	≈ 0.0987	≈ 0.0902	$0.5\overline{63}$
$\langle \pi_{nrc}, \pi_{rc} \rangle$	≈ 0.1974	≈ 0.1504	$0.\overline{24}$
$\langle \pi_{nrc}, \pi_{nrc} \rangle$	≈ 0.0987	≈ 0.6692	$0.\overline{96}$

Table 24: Results for L_{Cag}. As we did for preference functions, we chose to interpret the anomalous cases from P_S^* as giving out a zero probability on the message that they predict.

5.4.4 L_{Cag}

Stories of dogwhistles all have in common the fact that the dogwhistle ends up being discovered by the outgroup (otherwise there would be no reporting on them). Our model as it is envisions listeners as being extremely passive in the dogwhistle game, they are standard cooperative listeners that will interpret political messages standardly, as if all of it was standard coooperative communication. Even then, because of the double nature of the message being transmitted, we have seen that the interpretation of messages by the audience is no simple matter, trying to communicate two messages at once, may they be of different nature, will lead to uncertainty in the interpretation of those messages. While L_1 might be adequate to describe a part of the audience, it surely is not for those listeners that exhibit more political savvy or just generally keep themselves more informed, or listeners that are just inclined to think that they are not partaking in a cooperative interaction.

Enter the *cagey listener* L_{Cag}. Whereas L_1 had as an internal speaker model a S_{Reg} speaker, L_{Cag} has a S_{Dup} speaker in its place. A standard implementation of L_{Cag} then seeks to interpret an ordered set of worlds and an ordered set of personae from the message that they are exposed to. As shown in Figure 5, this leads to the initial $L_1(w|m)$ and $L_1(\pi|m)$ from L_1 to be replaced with their counterparts $L_{Cag}(\langle w, w' \rangle | m)$ and $L_{Cag}(\langle \pi, \pi' \rangle | m)$. Using these interpretation functions will lead us to have the results presented in Table 24, where the correct inferences are made when hearing the dogwhistle m_{dw}. There can however be more to this. As we have seen in the previous section, it makes sense to think of the speaker in a dogwhistle context to act according to preference functions, because of the non-cooperative nature of the exchange. The actual goal of a non-cooperative listener, in that context, would not simply be to infer the intended *content* of the message chosen by the speaker, it would be to infer the speaker's *preferences*.

$L_{Cag}(\langle w, w' \rangle, \nu^* \mid m)$	m_{aa}	m_{cc}	m_{dw}
$\langle w_{aa}, w_{aa} \rangle, \nu^*$	≈ 0.2203	≈ 0.0371	≈ 0.0363
$\langle w_{aa}, w_{cc} \rangle, \nu^*$	≈ 0.1282	≈ 0.1345	≈ 0.3403
$\langle w_{cc}, w_{aa} \rangle, \nu^*$	≈ 0.0598	≈ 0.0742	≈ 0.0725
$\langle w_{cc}, w_{cc} \rangle, \nu^*$	≈ 0.0382	≈ 0.2463	≈ 0.0989
$\langle w_{aa}, w_{aa} \rangle, \nu^{*'}$	≈ 0.3453	≈ 0.0572	≈ 0.0492
$\langle w_{aa}, w_{cc} \rangle, \nu^{*'}$	≈ 0.0672	≈ 0.0763	≈ 0.1199
$\langle w_{cc}, w_{aa} \rangle, \nu^{*'}$	≈ 0.1007	≈ 0.1145	≈ 0.0985
$\langle w_{cc}, w_{cc} \rangle, \nu^{*'}$	≈ 0.0403	≈ 0.2599	≈ 0.1043
$L_{Cag}(\langle \pi, \pi' \rangle, \mu^* \mid m)$			
$\langle \pi_{rc}, \pi_{rc} \rangle, \mu^*$	≈ 0.2333	≈ 0.0297	≈ 0.0332
$\langle \pi_{rc}, \pi_{nrc} \rangle, \mu^*$	≈ 0.0962	≈ 0.0807	≈ 0.3916
$\langle \pi_{nrc}, \pi_{rc} \rangle, \mu^*$	≈ 0.0598	≈ 0.0495	≈ 0.0875
$\langle \pi_{nrc}, \pi_{nrc} \rangle, \mu^*$	≈ 0.0477	≈ 0.3271	≈ 0.0443
$\langle \pi_{rc}, \pi_{rc} \rangle, \mu^{*'}$	≈ 0.3616	≈ 0.0458	≈ 0.0467
$\langle \pi_{rc}, \pi_{nrc} \rangle, \mu^{*'}$	≈ 0.0504	≈ 0.0458	≈ 0.2283
$\langle \pi_{nrc}, \pi_{rc} \rangle, \mu^{*'}$	≈ 0.1007	≈ 0.0763	≈ 0.1216
$\langle \pi_{nrc}, \pi_{nrc} \rangle, \mu^{*'}$	≈ 0.0504	≈ 0.3452	≈ 0.0467

Table 25: Results for L^*_{Cag}, with priors over possible preference functions. As we did for preference functions, we chose to interpret the anomalous cases from P^*_S as giving out a zero probability on the message that they predict. Our typical dogwhistle interpretations are in red, we can see that they are the most likely interpretation when hearing m_{dw} (probabilities in blue).

Because the set of preferences is technically infinite, for this to make sense and be computable, we limit ourselves to sets of relevant preference functions M, N which would be determined by the broader context (and possibly indexed to personae). This updated version of the cagey listener, L^*_{Cag}, has priors over the possible preference functions and infers them along with inferring the semantic and social content of utterances, leading to the results in Table 25 when we place uniform priors over possible preference functions $\in$ M, N defined as either the preference functions in Table 23 or $\mu^{*'}, \nu^{*'}$, defined as the absence of preferences (for all messages m, $\mu^{*'}(m) = \nu^{*'}(m) = 1$). What we observe is that in case this listener is exposed to message m_{dw}, not only will it infer that the duplicitous interpretations $\langle w_{aa}, w_{cc} \rangle, \langle \pi_{rc}, \pi nrc \rangle$ are more likely than the others, it will also infer that it is a lot more likely that the preference functions of S_{Dup} are μ^*, ν^* rather than $\mu^{*'}, \nu^{*'}$.

5.5 CONCLUDING REMARKS

We have described at length and tested DWG, our model for the formal description of dogwhistle interactions. It has interesting predictions in its most sophisticated representations, that are in keeping with the concept of dogwhistles as we have presented it in the previous chapters. Nonetheless, it also has limitations. For example, it makes fairly odd predictions when it

comes to standard listeners, who might end up making absurd interpretations from messages due to the fact that each message now allows to access two different kinds of information. While we can cancel this effect with some tweaking of the model, there is no satisfactory solution to it as it exists.

One thing that dogwhistles have forced us to do in terms of modeling is rethink a lot of what RSA-based models take for granted by trying to apply their logic to a non-cooperative situation. We have seen that the use of preference functions over both personae and worlds, while it is unjustified for the kind of communicative events that RSA describes, is useful here to make sense of some properties of dogwhistle exchange. We have also seen that even though to some extent DWG are extensions to both RSA and SMG, they have profound differences with both that can lead to an entirely different game, one where accurate information transmission is no longer a goal, and where the inference of individual preferences becomes more relevant.

The complete computations for the examples in this chapter can be reproduced using the methods presented in Appendix B.

As underlined in this chapter, this implementation of the model allows us to describe one very specific flavor of dogwhistle, a flavor that has been already put forward by previous work, notably Henderson and McCready (2018). There is more to dogwhistles than this, however, Saul (2018a) underlines the application of the concept to non-political contexts, notably in cultural content, where the stakes are vastly different from what we find in political discourse and where the notion of (non-)cooperation becomes a lot less clear; in Saul (2018b) we are also presented with *Protean dogwhistles*, where a single message is interpreted in many different ways by many different groups, and not targeting one specific ingroup. The next chapter will focus on these cases, discussing extensions to the DWG model that can be made to attempt to describe these cases.

6

EXTENDING THE FORMAL MODEL

"I celebrate myself, and sing myself,
And what I assume you shall assume,
For every atom belonging to me as good belongs to you."

— Opening lines of Whitman's *Song of Myself*. He seems to assume that readers will share his assumptions. (Whitman, 1855)

What we have tried to do in Chapter 5 is construct a formal model to describe situations where a speaker tries to communicate several different messages at once. Messages that both differ in nature (*social meaning* and *content meaning*) and in content (communicating one thing to a listener and another thing to another listener). We focused on a very simple and often treated case of dogwhistle communication involving just two listeners and a situation of ideological conflict. Dogwhistle communication can however take many forms. In Saul (2018a), dogwhistle communication outside the realm of politics is mentioned, and in Saul (2019), we are presented the concept of *Protean dogwhistles*, where the number of interpretations for a given term can be many more than just two. In this part, we will describe those two phenomena and attempt to extend our model to present an account of them.

In the context of politics, it makes sense that a candidate would want to maximize the number of people voting for them/agreeing with them – telling each possible audience whatever can motivate their voting for the candidate – given that success is in large part measured in popular support. But the general idea behind dogwhistling, that is communicating several possibly conflictual messages at once, can be applied to more mundane contexts where political ideologies are not as salient and where popular support plays a lesser role in determining success.

In Saul (2018a), we are confronted with an interpretation of dogwhistles that goes beyond the realm of political speech. Dogwhistles are there construed as being any kind of signal that communicates several meanings, with one meaning being concealed to a part of the audience. The example put forward in Saul (2018a) is that of children's cartoons that reference cultural content that is obviously unavailable to children but is available to their parents in order the make the watching of the cartoon as a family more interesting to parents. Without even going into the issue of cultural references, we argue that there are many similar phenomena of hidden meaning in cultural content. Even though these cases differ vastly from political discourse from a philosophical point of view (the notion of *duplicity* is not necessarily as prominent here), we think that these phenomena can be formally described in very similar, if not identical terms to the political discourse phenomena we described thus far.

The cases that we will study in the rest of this section concern readers or listeners' interpretations of literature, poetry and songs, where the meaning(s) that they assign to lines and lyrics are influenced by aspects of their beliefs about the writer's personal identity and their intention. It is, of course, common for works of art to be interpreted in different ways by different audiences; however, in order to make the parallel with dogwhistles as clear as possible, we will focus on instances where the persona of the speaker (and the listener's beliefs about it) are what is driving variation in interpretation (and in production).

6.1 WALT WHITMAN

Our first example of dogwhistle-like language use and interpretation comes from the works of Walt Whitman. Whitman (1819-1892) was an American poet, whose works focused on nature and life in the United States after the Civil War. He is one of the major figures of American transcendentalism, a philosophical and artistic movement which has as core beliefs the inherent goodness of people and nature and the celebration of self-reliance and individualism. Thanks to his literary success, Whitman has also become a sort of gay (or queer) icon: there is evidence that he was romantically involved with men, and modern critics seem to agree that he refers to homosexual relationships in poems of his magnum opus *Leaves of Grass* (Champagne, 2008). This being said, it does not seem that he publicly discussed his sexual preferences, and this has given rise to ambiguity and uncertainty about his sexual identity for those who read his work (Schmidgall, 1997).

We argue that the ambiguity surrounding Whitman's persona is similar to the ambiguity surrounding political personae in the more standard dogwhistle cases, and, furthermore, we show that it similarly gives rise to different content interpretations of lines of Whitman's poetry. For example, in the foreword to Schmidgall (1997), several references are made to Walt Whitman's being forced to hide his sexuality, a fact which supposedly accounts for the more "difficult" passages in *Leaves of Grass*:

> *"Where his poetry is 'difficult', it is often because of the methods of furtiveness, calculated ambiguity, and closeted concealment with which Whitman cushioned (. . .) many of his most strikingly bold assertions about sex and sexuality in Leaves of Grass." p.xvi*
>
> *"The present study (. . .) is largely about those particular 'amens' that were uttered by the select few readers of Leaves of Grass who did not see smut in its pages but rather a thrilling invitation to self-recognition, fraternity and empowerment (. . .) the 'amens' of readers who responded to the 'secret and divine signs' of its gay subtext. These are the readers whom, I believe, Whitman was quite consciously addressing." p.xx*
>
> — Schmidgall (1997)

This analysis of Whitman's poetry evokes dogwhistling very clearly, and in a sense that is close to what we have seen in political speech. Concepts like "*comradeship*", "*adhesiveness*"[1] or "*manly attachment*" are ambiguous as to whether they also concern the "physical aspects of manly love" (a concern of John Symonds, poet, correspondent and biograph of Whitman's, cited in Champagne, 2008). In sum, there seems to be two different readings of Whitman's "*comradeship*", and the vast literature on the author comprises both approaches emphasizing the importance of Whitman's identity as a gay man in the analysis of his works and the view of his homoeroticism as being more akin to a spiritual approach to male friendship (see the "Comradeship" article in LeMaster, Kummings, and Whitman, 1998).

At the center of this dichotomy is the collection of poems *Calamus* in *Leaves of Grass*, which treat at length the issue of comradeship and manly attachment with sexual undertones[2]. We argue in the following that the processes at work in the artist's writing style are in fact similar to what is at work in the use of dogwhistles in political speech, and that the model that we have used to describe the latter can also be used to describe the former, although before doing so, it is important that a number of limitations of our approach be underlined.

The model we have presented in Chapter 5 can be used both as a speaker and a listener model, like all models inspired by RSA. In this case however, an issue arises: if it does not sound out-of-place to give a model for a contemporary reader of Whitman's poetry, including one that would focus on an idealized writer's intentions, we should be much more circumspect about using the same interpretations, both in terms of content and social meaning, to characterize the speaker/writer, who in the present case has been dead for a very long time. In all cases, the 'true' intentions of a speaker or author are beyond our knowledge, but this is even more the case in cases like the present, where the author is from a completely different era. In fact, as underlined in Champagne (2008), it is likely that an identity adjective like 'gay', often used to refer to Whitman himself, had no particular relevance then, and if it did, most likely had a very different meaning, with the persona of, say, a 'gay man', not being accessible in any way to the poet's contemporaries.

Nonetheless, this does not forbid us from describing a current-era listener using our model, nor should it stop us from thinking of listener interpretation in terms of retrieval of speaker intention, in case we agree that mechanisms of intention retrieval are in fact at the heart of the interpretation process – as is the case in most game-theoretic pragmatics frameworks. Whether the internal speaker that listeners envision does in fact correspond to the actual speaker or not is irrelevant and does not limit the capabilities of our listener model. For a more convincing case regarding the speaker model and non-political dogwhistles, see Section 6.2.

1 A term Whitman took up from phrenology, where it refers to an organ whose size was supposedly proportional to individuals' predisposition to "attachment". See Combe (1830), pp.237-8, reproduced here:
http://www.historyofphrenology.org.uk/system/adhesiveness.htm
It has to be mentioned that along with his beliefs regarding phrenology, Whitman is known for having held racist beliefs and made racist comments during his lifetime, an aspect of his work that ought to be studied more in-depth, see this piece:
https://www.wnyc.org/story/walt-whitman-turns-200-and-poet-harmony-holiday-calls-out-his-racism/

2 The title itself, *Calamus*, refers to a root whose shape is often likened to that of a penis.

6.1.1 *Dogwhistles in* Leaves of Grass

Let us now consider an example in Whitman's poetry that could be construed as ambiguous: the poem *Among the Multitude*, reproduced here in full:

> AMONG THE MULTITUDE
>
> Among the men and women the multitude,
> I perceive one picking me out by secret and divine signs,
> Acknowledging none else, not parent, wife, husband, brother, child, any nearer than I am,
> Some are baffled, but that one is not – that one knows me.
>
> Ah lover and perfect equal,
> I meant that you should discover me so by faint indirections,
> And I when I meet you mean to discover you by the like in you.

The poem is among the *Calamus* poems (number 41) and its lines have been assigned a number of different interpretations by different readers. We will first focus on one very specific example to show that it can be treated in the same way we treated dogwhistles. Consider line 5 of the poem:

(22) Ah lover and perfect equal

(22), in the context of the poem can refer to either a man or a woman, and both interpretations can be found among readers. Based on their belief that Whitman possesses a homosexual identity, critics like Schmidgall (1997) insist that this poem should be understood as a hidden calling to the *likes* of the poet (understood here to be fellow gay readers):

> *"Here is Whitman's ideal reader, alive to the 'secret and divine' hints and indirections that reveal the 'life behind the life', the sex life."*
>
> — Schmidgall (1997)

According to Schmidgall (1997), the expression *lover and perfect equal* should be interpreted within the set of individuals possessing the *male* property in the domain. Clues that could lead readers to think that the person is male instead of female lie on several things. According to Schmidgall (1997), the "secret and divine signs" as well as the "faint indirections" alluded to by the author point towards his subtle displaying of his sexuality to the crowd in such a manner that only a chosen few might correctly read his signals; the "faint indirections" are also found in the lover (the "like" in the last line refers then to the "indirections" of the line above). In other words, there are a number of "secret signs" that Whitman displays and

recognizes in a group of people which he then considers "lover[s]" and "perfect equal[s]"[3]. Notice that no explicit mention of the gender is made in the poem; however, in Schmidgall (1997), it seems obvious that the secrecy of the exchange that is mentioned is a hint towards the thought that the recipient is male.

It is however mentioned in Schmidgall (1997) that, although he believes it to be the 'correct' one, the *male* interpretation of (22) is not the only one that exists:

> *"I am frankly suspicious of all these 'civilian' critics who seem eager to pronounce the lubricious matter a dead end and discussion of it exhausted. [...] For these critics are precisely the ones most likely to 'fumble at the lock' on the door of truth about Whitman's sex."*
>
> — Schmidgall (1997)

An example of another interpretation by such a "civilian critic" is shown below, where, in this case, the blog author clearly interprets *lover* as being female:

> *"The joy comes with the sudden realization that he is not alone. In fact, he has found the perfect other, in a happenstance that not of his own making. Perhaps in some sense, he has entered into a higher realm of consciousness, where he has encountered the divine or the feminine divine."*
>
> — Excerpt from a blog post on the blog Love 2013 AD[4]. (W. Diane Van Zwol, 2014)

Another example of a *female* interpretation of *lover and perfect equal* is shown below, and note that this writer also later goes on to refer back to the *lover* using the pronoun *she*.

> *"After reading this poem a couple of times, one of the first things that caught my attention was sentence number two. 'I perceive one picking me out by secret and divine signs,' seemed innocent and young, as if the author was relating to his childhood days when he saw the girl that took his breath away, but couldn't tell anyone because girls have cooties."*
>
> — Excerpt from a blog post on the blog The Difficulty of the Long Twentieth Century. (xpo154, 2010)

In summary, the fact that Whitman leaves the door open on the gender of the recipient of the secret and divine signs allows a reader who believes Whitman to be heterosexual (either because this is their understanding of the historical facts or because they see heterosexuality as the 'default' sexual orientation) to interpret the poem to refer to a woman, as do most poems written in this fashion [5]. Conversely, people that are aware, in one way or another, of Whitman's sexuality, will read in his verse that the "lover" may in fact be male.

3 This also sounds like a form of dogwhistling in a sense. It is important that although this work is about language, style and persona-building are not merely language-specific and can be done in many ways. So can the concept of dogwhistling be applied to gestures, attitude, etc.

4 Note in that case that the line in question has been wrongfully transcripted as "As lover and perfect equal". This does not change the analysis we make of it, but it does change the semantics of the sentence quite a bit, in the case of the alternate transcription, perhaps the "one" of line 2 is a more interesting example for ambiguous reference.

5 See for example the fairly similar, though explicit regarding gender, *À une passante*, by Charles Baudelaire.

In other words, the meaning of the poem depends on a priori beliefs regarding Whitman's identity as either[6]:

- A gay poet
- A straight poet

6.1.2 *An attempt at a formalization*

As said in Chapter 1, a formal model can allow us to see the underlying commonalities between phenomena that are different at first glance. In case we agree that political dogwhistling is not unlike the examples of cultural content presented in Saul (2018a), then an appropriate model for the description of political dogwhistle communication should with little adjustment also be usable to describe such instances of non-political dogwhistles. Let's therefore try to apply our model to the example of *Among the Multitude*. From what we have said, we can infer two different readings for the poem:

1. Gay[7] romance
2. Straight romance

As was the case with "inner cities", we can think of alternatives to (22) that would be unambiguous regarding the gender of their referent. We have to keep in mind, obviously, that poetry does not necessarily abide by the same rules as regular language and that many aesthetic parameters are taken into account when choosing to write lines as they are written. We choose here to ignore these parameters and greatly simplify the situation when proposing these alternatives to (22):

(23) a. Ah gallant and perfect equal

b. Ah donna and perfect equal

With (23a) an unambiguous male reference and (23b) being an unambiguous female reference, both keeping some connotations of romantic partnership.

It is then rather straightforward to define terms symmetrically to what we had in Section 5.4.

Let Γ be a DWG: $\langle \{S, \{L_{1_g}, L_{1_s}\}\}, L_{0_{g/s}}, W, M, \Pi, \textsc{Lex}, \textsc{Soc}, \Pi\text{-}\textsc{Lex}, \Delta\text{-}\textsc{Soc}, Pr_W, Pr_\Pi \rangle$.

Once again, temperature parameters $\alpha, \alpha', \beta, \beta'$ are all set to 1.

- S, the speaker, would be Walt Whitman. As stated earlier, making a speaker model of Walt Whitman here makes little sense, but the framework calls for a speaker in any case.

6 Note also that given the nature of the text, it could very well be that the character behind the voice of the poem is a creation of the poet that could, for example, be a woman. We will here assume that the voice of the poem is in fact the voice of the poet.

7 For lack of a better word.

- $L1_g, L1_s$ are two pragmatics listeners, respectively a *gay friendly reader* and a *straight default reader*
- $L0_{g/s} : M \to \Delta(W), \Delta(\Pi)$ are Literal Listeners corresponding to the priors found below, one for each listener.
- $W = \{w_m, w_f\}$ is a set of possible states of the world, respectively one where the individual referred to by S using (22) is male and one where they are female.
- $M = \{(22), (23a), (23b)\}$ is the set of relevant possible messages available to the speaker.
- $\Pi = \{\pi_{gp}, \pi_{sp}\}$ where π_{gp} is the persona of a *gay poet* and π_{sp} the persona of a *straight poet*.
- $\textsc{Lex} = \{[\![\cdot]\!]_{gp}, [\![\cdot]\!]_{sp}\}, \textsc{Soc} = \{[\cdot]_g, [\cdot]_s\}$ we have one interpretation function per persona and one indexation function per listener.
- We set Δ-Soc as follows:

$$\Delta\text{-Soc}([\cdot]_g) = p$$
$$\Delta\text{-Soc}([\cdot]_s) = 1 - p$$

  With $p = 1$ for $L1_g$ and $p = 0$ for $L1_s$. As in Section 5.4.2, we assume that each listener only has access to one indexation function.

- We set Π-Lex as follows, identically for both listeners:

$$\Pi\text{-Lex}(\pi_{gp}) = \begin{cases} [\![\cdot]\!]_{gp} : 1 \\ [\![\cdot]\!]_{sp} : 0 \end{cases}$$
$$\Pi\text{-Lex}(\pi_{sp}) = \begin{cases} [\![\cdot]\!]_{gp} : 0 \\ [\![\cdot]\!]_{sp} : 1 \end{cases}$$

  This time we do not assume that the *straight default reader* $L1_s$ merely ignores that words referring to romantic partners, when uttered by people constructing a male gay persona, have male referents. While it was justified in the political dogwhistle scenario, it would now just be a silly idea. This hints at a very important difference between this case in the one in Section 5.4: the ignorance of one of the listeners acts on a different level here, they do not ignore the fact that the referent of the word "lover" can be male, what they

[m]	$[\cdot]_g$	$[\cdot]_s$
(22)	$\{\pi_{sp}, \pi_{gp}\}$	$\{\pi_{sp}\}$
(23a)	$\{\pi_{gp}\}$	$\{\pi_{gp}\}$
(23b)	$\{\pi_{sp}\}$	$\{\pi_{sp}\}$

Table 26: Definition of the indexation functions $[\cdot] \in$ Soc. We assume that an overt reference to a male romantic partner will lead to interpreting that the poet displays a *gay poet* persona.

ignore is this time at the level of the indexation function, where they will only recognize a gay persona if it is explicitly mentioned, assuming a default straight persona. We will see how that changes the results that our model produces.

- The priors over worlds Pr_W are uniform. The priors over personae Pr_Π are uniform as well.

The indexation and interpretation functions for the messages in M are shown in Tables 26 and 27, leading to the L_0 interpretations in Tables 28 and 29. A short point on the interpretation functions $[\![\cdot]\!] \in$ Lex: in this case, it is maybe not so clear that they should be interpreted like group-based dialects. In fact, it is obvious that a word like "lover" or "partner" can have both male and female referents no matter which group you find yourself in. This is more a case of the referent of the word being set once the identity of the speaker and their relationship to that referent are known. Compare for example the following:

(24) My lover loves her lover.

These two instances of "lover" can reference two people of different genders, regardless of the group to which the speaker belongs. So for example, $[\![\cdot]\!]_{gp}$ in Table 27 is to be read not as "a person using that interpretation function would use the word 'lover' to refer to a male individual", but rather as "the referent of this instance of 'lover' is to be interpreted as male in case the speaker is using this interpretation function". This is different from what we had before because we go *beyond semantics* and fill in some of the blanks necessary for the truth-value of the sentence to be computed. In fact, and as was alluded to in Chapter 2, the issue of reference is more an issue of *pragmatics* than it is of *semantics*. That being said, we can argue in favor of such an approach to the interpretation function by assuming that what we are interested in here is not the meaning of the *words* "lover", "gallant" and "donna", but of the *lines* (22), (23a) and (23b) in the context of the poem. The lines make it evident that the referent of the word is thought to be a potential "lover" to the poet and none else, hence the restriction of the reference that we put forward in Table 27. This sentence appears in a context where it is made obvious that the "lover" relationship is experienced by the speaker, a *straight default* reader will assume that the sentence can only be true if the genders of the two lovers are distinct, while a *gay friendly* reader will not. But we are cheating a little bit here.

$[\![m]\!]$	$[\![\cdot]\!]_{gp}$	$[\![\cdot]\!]_{sp}$
(22)	$\{w_m\}$	$\{w_f\}$
(23a)	$\{w_m\}$	$\{w_m\}$
(23b)	$\{w_f\}$	$\{w_f\}$

Table 27: Definition of the interpretation functions $[\![\cdot]\!] \in$ LEX.

$P_{L_0}(w\|m)$	w_m	w_f
(22)	0.5	0.5
(23a)	1	0
(23b)	0	1
$P(\pi\|m)$	π_{gp}	π_{sp}
(22)	0.5	0.5
(23a)	1	0
(23b)	0	1

Table 28: Listeral listener results for the gay friendly reader L_{gf}.

$P_{L_0}(w\|m)$	w_m	w_f
(22)	0	1
(23a)	1	0
(23b)	0	1
$P(\pi\|m)$	π_{gp}	π_{sp}
(22)	0	1
(23a)	1	0
(23b)	0	1

Table 29: Listeral listener results for the straight default reader L_{sd}.

$P_{L_1}(w\|m)$	w_m	w_f
(22)	0.5	0.5
(23a)	0.75	0.25
(23b)	0.25	0.75
$P(\pi\|m)$	π_{gp}	π_{sp}
(22)	0.5	0.5
(23a)	0.75	0.25
(23b)	0.25	0.75

Table 30: Pragmatic listener results for the gay friendly reader L_{gf}.

$P_{L_1}(w\|m)$	w_m	w_f
(22)	≈ 0.286	≈ 0.714
(23a)	0.8	0.2
(23b)	≈ 0.286	≈ 0.714
$P(\pi\|m)$	π_{gp}	π_{sp}
(22)	≈ 0.286	≈ 0.714
(23a)	0.8	0.2
(23b)	≈ 0.286	≈ 0.714

Table 31: Pragmatic listener results for the straight default reader L_{sd}.

With both listeners assuming a S_{Reg} speaker, we end up with the results found in Tables 30 and 31. We observe that the interpretation of "lover" for *gay friendly* readers stays uncertain, stuck at an equal probability of the reference being male or female, while the alternatives lead to clearer interpretations. If we look at *straight default* readers, the interpretations are less uncertain, with "lover" most likely referring to a female individual. We also see that, in keeping with our interpretation and indexation functions, (22) and (23b) are perfect synonyms for the *straight default* reader and will lead to the same results. Importantly, (22) rarely leads to a w_m interpretation for the *straight default* reader, or at least a lot less than for the *gay friendly* reader, which is the key interpretation point that this was about making.

Let's take a step back for a moment. The whole system is fairly complicated, and there might be an intuitive approach to the question that at first seems more reasonable: why not have one single interpretation function (as stated earlier, it is unlikely that the semantics of "lover" truly differ between groups) and just change the listener priors about the situation to have the situation that we are looking for? First of all, we must try to assess what "changing the listener priors" actually means. Several things in our model can be identified as "listener priors". In a typical RSA setting, the priors are over the possible worlds, with an assumption that listeners can have some knowledge about the state of the world before receiving any kind of signal from the speaker. Changing the priors over possible worlds basically means that the listener has information about the situation that the speaker is discussing. It seems strange in the

$[\![m]\!]$	$[\![\cdot]\!]_\iota$
(22)	$\{w_m, w_f\}$
(23a)	$\{w_m\}$
(23b)	$\{w_f\}$

Table 32: Definition of the interpretation function $[\![\cdot]\!]_\iota$.

$P_{L_1}(w\|m)$	w_m	w_f
(22)	0.5	0.5
(23a)	≈ 0.768	≈ 0.232
(23b)	≈ 0.266	≈ 0.734
$P(\pi\|m)$	π_{gp}	π_{sp}
(22)	≈ 0.641	≈ 0.359
(23a)	≈ 0.696	≈ 0.304
(23b)	≈ 0.203	≈ 0.797

Table 33: Pragmatic listener results for the gay friendly reader in the case where only $[\![\cdot]\!]_\iota$ is taken into account and priors over personae are modified.

present case to assume that, since the situation is purely fictional, the listener is not a witness to it nor do they have a prior information from another source. It therefore seems uncalled for to change the priors over worlds for the listeners here. Which leaves us with priors over personae, knowledge about different indexation functions (Δ-Soc) and association between persona and interpretation function (Π-Lex). The one that it makes most sense to modify is probably the priors over personae; in fact, as we have set them, they reflect a reader that has no information about the identity of the speaker, which seems unlikely in many cases. Let's therefore try the following: we will only use the $[\cdot]_g$ indexation function and the interpretation function $[\![\cdot]\!]_\iota$ in Table 32 for all cases, meaning that for both listeners, $\Delta\text{-Soc}([\cdot]_g) = 1$ and $\pi \in \Pi$, $\Pi\text{-Lex}(\pi)$ leads to $[\![\cdot]\!]_\iota$ with probability 1. We however change the priors over personae such that they now are as follows:

$$Pr_\Pi(\pi_{gp}) = p$$
$$Pr_\Pi(\pi_{sp}) = 1 - p$$

With $p = 0.9$ for L_{gf} and $p = 0.1$ for L_{sd}.

As can be seen in the pragmatic listener results for this in Tables 33 and 34, what we obtain in that case is that the uncertainty relative to the gender identity of the referent of "lover" is generalized to both listeners, but that the persona interpretation $P(\pi|m)$ is aligned with the priors. We lose a lot of the difference in content interpretation because the way that

$P_{L_1}(w\|m)$	w_m	w_f
(22)	0.5	0.5
(23a)	≈ 0.734	≈ 0.266
(23b)	≈ 0.232	≈ 0.768
$P(\pi\|m)$	π_{gp}	π_{sp}
(22)	≈ 0.359	≈ 0.641
(23a)	≈ 0.797	≈ 0.203
(23b)	≈ 0.304	≈ 0.696

Table 34: Pragmatic listener results for the straight default reader in the case where only $[\![\cdot]\!]_\iota$ is taken into account and priors over personae are modified.

the persona of the speaker can influence the content interpretation in this model happens mostly through differentiated lexica. If we drop the difference in lexica, then we lose most of the interpretative variation, and in the present case, the least precise formulation, (22) leads to absolute content uncertainty for both listeners, since there always is a more informative alternative available. Using exclusively priors modulation to reach the result that we want would require us to modify the world priors as well, which is equivalent to assuming that the listener knows in advance what the speaker is likely to communicate, which in this situation does not sound plausible. The complexity of the model used here is justified by the fact that we want a variation in content interpretation that is based on speaker persona, not simply on a priori knowledge about the world.

6.1.3 *Speaker model*

The listener models above capture how readers' beliefs about Walt Whitman's sexual identity can influence the interpretation of the language in his poems. However, listener models are only one side of the coin, and we might also like to develop a speaker model of Whitman. As stated earlier, this is not a trivial matter, and it may in fact make little sense to do so. Let's try to make sense of what that would entail.

Firstly, it has been observed that, notably, it might make little sense for Whitman himself and his contemporaries to identify as a "gay poet" or a "straight poet", since our modern "gay" and "straight" identities might very well not have had any kind of resonance at the time, in that place. This does not affect our listener models, since the modern listeners in our examples can attribute these identities to Whitman. However, Whitman himself may have considered his sexual persona to be more along the lines of what we now call *queer* (Champagne, 2008), or something else entirely.

Secondly, even if the only personae at play are *queer* and *non-queer*, there are many modelling options available. One possiblity is that Whitman himself *is* actually queer or *is not* queer, and wants to communicate this. In which case, he would have a single message w_m or w_f, and

then choose one of (22), (23a) or (23b) to write. Whitman ended up choosing the possibly ambiguous *lover*, so, according to our model, two things could be going on: the first one is that Whitman assumes that his audience is uniformly composed of something like straight default readers. In this case, the word *lover* will most probably be interpreted as picking out women.

An alternative is that what we observe is an instance of what we called the *duplicitous speaker*: Whitman's content meaning is a pair $\langle w_m, w_f \rangle$, and he aims for gay friendly readers to interpret w_m and straight default readers to interpret w_f. This use of the ambiguous expression could be the result of something that ressembles political dogwhistling, where, because of the stigma on homosexuality, Whitman feels like he needs to hide the male lover interpretation from outgroup listeners who may be hostile to his sexual identity[8]. Yet another possiblity is that Whitman is not motivated by duplicity; rather, consistent with his transcendentalist philosophy, the ambiguous expression is used to genuinely allow different kinds of listeners to arrive at different interpretations, depending on their subjective experiences.

In the spirit of the transcendental (not actually) duplicitous speaker is that Whitman may actually wish to communicate both w_m and w_f *at the same time* to any given reader. In which case, he is hoping for his audience to be consituted from what we called *cagey listeners*: readers who acknowledge both readings as acceptable according to the audience as opposed to those who wish to see one clear meaning in a string of text[9].

In any case, we have reasons to think that the scenario presented here in the interpretation of *Among the Multitude* is more complex than we made it look like. For example, we can assess that there exists at least one other possible reading of the poem which involves little to no romantic dimension. Such a reading can be found in Athenot (2009). This analysis voluntarily omits references to erotic content on the basis that in reading Whitman, resorting to what is explicitly said in the poem is a "trap", and that the poem should be read instead in the broader motifs of the celebrations of a solipsist self and of democracy:

> *"The democratic pose (perfect equal) and erotic innuendos led critics – who see in this a scene of flirting in a big city – to largely fall into the trap of the explicit. [...] In this encounter between the half-wanderer half-onlooker poet and the individual moving among the multitude, the issue of democratic fusion and narcissistic solipsism – the two poles between which Whitman's poetry constantly and recklessly oscillates – is again at stake. With democracy meeting modernity, it is the reader that Whitman imbues with the mission of cultivating the possibilities of poetry by opening it to its many meanings.[10]"*

8 See the documentary *The Celluloid Closet* (Epstein and Friedman, 1995) for many examples of strategic signalling in these kinds of circumstances.

9 We should note in that case that given the nature of poetry, this could concern a lot more readers than the cagey listener type would concern listeners in a political context.

10 "La pose démocratique (perfect equal) et les sous-entendus érotiques ont conduit les critiques, qui y voient une scène de drague dans une grande ville, à tomber en masse dans le piège de l'explicite. [...] Dans cette rencontre entre le locuteur mi-flâneur mi-badaud et l'individu qui se meut au cœur de la multitude se joue encore et toujours la question de la fusion démocratique et du solipsisme narcissique, les deux pôles entre lesquels oscille constamment et non sans goût du risque l'ensemble de la poésie whitmanienne. Et, geste où la modernité rejoint

— Athenot (2009)

This reading does underline that there may be an erotic reading of the poem, but that it is not the only reading available, nor that it even is the most interesting with regards to the rest of Whitman's works. This is remarkably far from the reading presented in Schmidgall (1997) that was mentioned earlier, which emphasizes the importance of sexuality in reading Whitman's poetry. This is also different from a reading of the situation where the speaker assumes that their listeners are all *cagey listeners*, in the sense that what is understood from the line is not actually that the referent might be either male or female (or anything in between), but rather that the absence of any clear specification of gender should lead us to think that resolving that reference is besides the point of the poem and should be ignored, a fact that might already be hinted at by a previous line:

(25) Acknowledging none else, not parent, wife, husband, brother, child, any nearer than I am

In that line, the fact that both the "wife" and "husband" relation that the individual in the crowd could have with individuals around them are explicitly named does seem to underline that the individual might be either male or female, or really anything else, the only thing that matters being that they are individuals among other individuals, and transforming the poem from a romantic encounter to a recognition of the "like" in the other, a fairly different interpretation. Poetry is complicated.

There would be many ways to implement a speaker model of Whitman using what we know and have said, but given the distance in time that separates us from him, no implementation would seem to be accurate or truly justified, and the matter is perhaps better left undiscussed; not only do we not have access to the speaker's intentions in that case, but the speaker is from a context so different from ours that any abstraction or attempt at describing their internal processes – especially regarding the kinds of personae they might wish to display – would seem pointless. We can however try to do so using a more contemporary example.

6.2 SUFJAN STEVENS

Since Whitman died over 100 years ago, we think it is impossible to go much further into his speaker model. However, the interpretative variation observed with Whitman also exists with more modern artists, for example, singer songwriter Sufjan Stevens (b. 1975), whose lyrics illustrate a similar phenomenon. In this case, the juxtaposition of Christian terminology and references along with homoerotic narratives have led listeners to make both interpretations as a function of whether they think his persona is closer to that of a *queer songwriter* or a *Christian songwriter*. As with all cases involving literary texts, using the tools of formal pragmatics will

la démocratie, c'est le lecteur que Whitman investit de la mission de cultiver les possibles de la poésie en l'ouvrant à ses sens multiples."

seem insufficient, or inappropriate, but given the time proximity of this instance of language, it seems easier to justify than it was with Whitman.

6.2.1 *Dogwhistles in* Seven Swans

The song *To Be Alone with You*, on Stevens' 2004 album *Seven Swans* illustrates this very well. The genius.com page for the song[11] can be used to see that there seem to be several different interpretations of the lyrics. The title itself is used ambiguously in the first hook of the song, where it is unclear whether "you" refers to an unnamed male individual or more specifically to Jesus Christ. The entire song allows for both interpretations. For example, the closing lines of the song:

(26) I've never known a man who loved me

has interpreted in both ways in the context of the song, as underlined in the genius.com annotation associated with it:

> *"It may not only be Jesus but also perhaps it is maybe about Men In Love I mean it's not that much of a stretch. It could be about either."*
>
> — genius.com (https://genius.com/3385593)

Most lines of the song are annotated with religious readings, and then the annotations are commented upon underlining the homosexual romance reading. The general understanding seems to be in this case that Stevens is using both Christian and queer themes to illustrate his chosen persona of a queer Christian songwriter. Following what we have said on Whitman, we can say that the speaker model for Stevens here, given the variety in the interpretation of his songs[12], can be explained in three ways: he may be trying to convey only one specific message and one specific persona, perhaps suspecting his audience is homogeneous, and the variety of interpretation results from a miscalculation on his part; he is a *duplicitous* speaker, wishing to give each group in the audience their preferred message; he actively seeks to communicate both identities (and thus both messages) *at the same time* to all listeners, thus constructing his identity of Christian queer songwriter. Given the several, often subtle, religious references in his texts as well as his active involvement in the LGBT+ community, it seems likely that this third option is the most accurate.

In DWG terms, Sufjan Stevens is a *duplicitous* speaker, but one that seeks not to deceive the audience. His goal seems to be not merely to 'trick' Christian people into interpreting a spiritual meaning, nor does it seem to be to 'trick' LGBTQ+ people into interpreting instances of same-sex romances in his works. Because Stevens lives in the 21st century, it also seems

11 genius.com is a website where users can provide song lyrics with their interpretation, the page in question can be found here: https://genius.com/Sufjan-stevens-to-be-alone-with-you-lyrics.

12 Other good examples include *Futile Devices*, *John My Beloved* or *All of Me Wants All of You*.

unlikely that he would seek to *hide* his queer identity, like we could suspect an author like Whitman might have wanted to. A more interesting reading of this is that Stevens seeks rather to communicate a *double identity*, as both a *queer songwriter* and a *Christian songwriter*. This intuitively seems compatible with DWG as we have presented them, especially if we take into account the notion of a *cagey listener*, since this family of listener specifically interprets pairs of messages and personae. Let's briefly try to describe a formal implementation of the situation that comes with the interpretation of (27).

6.2.2 *An attempt at a formalization*

Let's consider the following sentence and two possible alternatives:

(27) a. To be alone with you

b. To be alone with my lover

c. To be alone with Jesus

Let Γ be a DWG: $\langle \{S, \{L_{1_r}, L_{1_c}\}\}, L_{0_{q/c}}, W, M, \Pi, \textsc{Lex}, \textsc{Soc}, \Pi\text{-}\textsc{Lex}, \Delta\text{-}\textsc{Soc}, Pr_W, Pr_\Pi \rangle$.
Once again, temperature parameters $\alpha, \alpha', \beta, \beta'$ are all set to 1.

- S, the speaker, would be Sufjan Stevens here.
- L_{1_r}, L_{1_c} are two pragmatics listeners, respectively a *romantic listener* and a *Christian listener*.
- $L_{0_{r/c}} : M \to \Delta(W), \Delta(\Pi)$ are Literal Listeners corresponding to the priors found below, one for each listener.
- $W = \{w_m, w_{jc}\}$ is a set of possible states of the world, respectively one where S means 'to be alone with his lover' and one where S means 'to be alone with Jesus Christ'.
- $M = \{(27a), (27b), (27c)\}$ is the set of relevant possible messages available to the speaker.
- $\Pi = \{\pi_{qs}, \pi_{cs}\}$ where π_{qs} is the persona of a *queer songwriter* and π_{cs} the persona of a *Christian songwriter*.
- $\textsc{Lex} = \{[\![\cdot]\!]_{qs}, [\![\cdot]\!]_{cs}\}, \textsc{Soc} = \{[\cdot]_q, [\cdot]_c\}$ we have one interpretation function per persona and one indexation function per listener.
- We set Δ-Soc as follows:

$$\Delta\text{-}\textsc{Soc}([\cdot]_r) = p$$
$$\Delta\text{-}\textsc{Soc}([\cdot]_c) = 1 - p$$

[m]	$[\cdot]_r$	$[\cdot]_c$
(27a)	$\{\pi_{qs}\}$	$\{\pi_{cs}, \pi_{qs}\}$
(27b)	$\{\pi_{qs}\}$	$\{\pi_{qs}\}$
(27c)	$\{\pi_{cs}\}$	$\{\pi_{cs}\}$

Table 35: Definition of the indexation functions $[\cdot] \in$ Soc. We assume that overt references to Jesus Christ are necessarily linked to a Christian persona. The 'dogwhistled' reference to Jesus Christ is judged to be unlikely to be recongized by non-Christians.

With $p = 1$ for L_{1_r} and $p = 0$ for L_{1_c}. Each listener only has access to one interpretation function.

- We set Π-Lex as follows, identically for both listeners:

$$\Pi\text{-Lex}(\pi_{qs}) = \begin{cases} [\![\cdot]\!]_{qs} : 1 \\ [\![\cdot]\!]_{cs} : 0 \end{cases}$$

$$\Pi\text{-Lex}(\pi_{cs}) = \begin{cases} [\![\cdot]\!]_{qs} : 0 \\ [\![\cdot]\!]_{cs} : 1 \end{cases}$$

- The priors over worlds Pr_W are uniform. The priors over personae Pr_Π are uniform as well.

The indexation and interpretation functions for the messages in M are shown in Tables 35 and 36, leading to the L_0 interpretations in Tables 37 and 38 and the L_1 interpretations in Tables 39 and 40. With the way we have set the parameters, the Christian listener L_c appears to be a part of the ingroup, in dogwhistle terms, whereas the romantic listener L_r would be a member of the outgroup, unaware of the Christian meaning of some of the verse. Interestingly, we faced the opposite situation with Whitman, where we assumed that readers identifiying as *gay friendly* were the group that was more able to recognize the "secret and divine signs". The distribution of ingroupness and outgroupness is determined by the social/political/cultural context in which a message is uttered. If Stevens had written at the time of, say, Whitman, it might have been that the situation would have been reversed, with (27a) being a dogwhistle term indicating Stevens' romantic involvement, and not his spiritual involvement. To translate this situation, we could have gone further and set $[(27a)]_c = \{\pi_{cs}\}$ to reflect the fact that both communities, romantic listeners and Christian listeners, can be thought of as being the ingroup. Given the time of writing of the song, however, it did not seem unreasonable to set the indexation function the way we have.

$[\![\mathrm{m}]\!]$	$[\![\cdot]\!]_{\mathrm{qs}}$	$[\![\cdot]\!]_{\mathrm{cs}}$
(27a)	$\{w_{\mathrm{m}}\}$	$\{w_{\mathrm{jc}}\}$
(27b)	$\{w_{\mathrm{m}}\}$	$\{w_{\mathrm{m}}\}$
(27c)	$\{w_{\mathrm{jc}}\}$	$\{w_{\mathrm{jc}}\}$

Table 36: Definition of the interpretation functions $[\![\cdot]\!] \in \textsc{Lex}$.

	$P_{L_0}(w|\mathrm{m})$	w_{m}	w_{jc}
(27a)		1	0
(27b)		1	0
(27c)		0	1
	$P(\pi|\mathrm{m})$	π_{qs}	π_{cs}
(27a)		1	0
(27b)		1	0
(27c)		0	1

Table 37: Listeral listener results for the romantic listener L_{r}.

	$P_{L_0}(w|\mathrm{m})$	w_{m}	w_{jc}
(27a)		0.5	0.5
(27b)		1	0
(27c)		0	1
	$P(\pi|\mathrm{m})$	π_{qs}	π_{cs}
(27a)		0.5	0.5
(27b)		1	0
(27c)		0	1

Table 38: Listeral listener results for the Christian listener L_{c}.

$P_{L_1}(w|\mathrm{m})$	w_{m}	w_{jc}
(27a)	≈ 0.714	≈ 0.286
(27b)	≈ 0.714	≈ 0.286
(27c)	0.2	0.8
$P(\pi|\mathrm{m})$	π_{qs}	π_{cs}
(27a)	≈ 0.714	≈ 0.286
(27b)	≈ 0.714	≈ 0.286
(27c)	0.2	0.8

Table 39: Pragmatic listener results for the romantic listener L_{r}.

$P_{L_1}(w\|m)$	w_m	w_{jc}
(27a)	0.5	0.5
(27b)	0.75	0.25
(27c)	0.25	0.75
$P(\pi\|m)$	π_{qs}	π_{cs}
(27a)	0.5	0.5
(27b)	0.75	0.25
(27c)	0.25	0.75

Table 40: Pragmatic listener results for the Christian listener L_c.

With the parameters we have set, we observe a difference between both listeners that reflects some of our intuitions. Typically, a Christian listener L_c will generally have a higher probability of interpreting (27a) as referring to Jesus Christ than a romantic listener L_r would. When it comes to social meaning interpretation, L_c is more likely to identify someone who uses (27a) as being a Christian songwriter π_{cs}, whereas L_r is more likely to identify them as a queer songwriter π_{qs}. As mentioned before, however, in the case of Stevens, these pragmatic listeners are not necessarily what we're most interested in. In Table 41, we explore what a model involving the two literal listeners in Tables 37 and 38 gives as a result when we focus on the duplicitous speaker that we argued above might be a better representation of Stevens here.

As was the case in Chapter 5, we see that an anomalous case arises, which is the situation that is the exact opposite of the dogwhistled content we attribute to Stevens: wanting to communicate a *Christian songwriter* identity and a w_{jc} interpretation of the message to romantic listeners, while at the same time wanting to communicate a *queer songwriter* identity and a w_m interpretation of the message to Christian listeners. What this says about our model is basically that the way we have defined our set of messages M, there exists no message that could possibly convey that meaning. What we also observe is that the situation that we wanted to illustrate here, $\langle w_m, w_{jc}\rangle, \langle \pi_{qs}, \pi_{cs}\rangle$, will invariably lead to the use of (27a) over the alternatives. Like we did in Section 5.4.3.1, we can add preferences to the mix. Without preferences, we have the following results for $\mathcal{P}(m)$:

$$\mathcal{P}(27a) \approx 0.3791$$

$$\mathcal{P}(27b) \approx 0.2375$$

$$\mathcal{P}(27c) \approx 0.3208$$

Unlike the case in Section 5.4.3.1, we already observe a slight preference for the dogwhistle version, (27a). If we add a notion of preferences to our liking, we can reinforce this tendency.

$P^*_S(m\|\langle w, w'\rangle\langle \pi, \pi'\rangle)$	(27a)	(27b)	(27c)
$\langle w_m, w_m\rangle\langle \pi_{qs}, \pi_{qs}\rangle$	$0.\overline{3}$	$0.\overline{6}$	0
$\langle w_m, w_m\rangle\langle \pi_{qs}, \pi_{cs}\rangle$	0.5	0.5	0
$\langle w_m, w_m\rangle\langle \pi_{cs}, \pi_{qs}\rangle$	$0.\overline{3}$	$0.\overline{6}$	0
$\langle w_m, w_m\rangle\langle \pi_{cs}, \pi_{cs}\rangle$	0.2	0.4	0.4
$\langle w_m, w_{jc}\rangle\langle \pi_{qs}, \pi_{qs}\rangle$	0.5	0.5	0
$\langle w_m, w_{jc}\rangle\langle \pi_{qs}, \pi_{cs}\rangle$	1	0	0
$\langle w_m, w_{jc}\rangle\langle \pi_{cs}, \pi_{qs}\rangle$	1	0	0
$\langle w_m, w_{jc}\rangle\langle \pi_{cs}, \pi_{cs}\rangle$	$0.\overline{3}$	0	$0.\overline{6}$
$\langle w_{jc}, w_m\rangle\langle \pi_{qs}, \pi_{qs}\rangle$	$0.\overline{3}$	$0.\overline{6}$	0
$\langle w_{jc}, w_m\rangle\langle \pi_{qs}, \pi_{cs}\rangle$	1	0	0
$\langle w_{jc}, w_m\rangle\langle \pi_{cs}, \pi_{qs}\rangle$	?	?	?
$\langle w_{jc}, w_m\rangle\langle \pi_{cs}, \pi_{cs}\rangle$	0	0	1
$\langle w_{jc}, w_{jc}\rangle\langle \pi_{qs}, \pi_{qs}\rangle$	0.2	0.4	0.4
$\langle w_{jc}, w_{jc}\rangle\langle \pi_{qs}, \pi_{cs}\rangle$	$0.\overline{3}$	0	$0.\overline{6}$
$\langle w_{jc}, w_{jc}\rangle\langle \pi_{cs}, \pi_{qs}\rangle$	0	0	1
$\langle w_{jc}, w_{jc}\rangle\langle \pi_{cs}, \pi_{cs}\rangle$	0	0	1

Table 41: Results for S_{Dup}. The preferred message for each situation is in blue, the situation that is closest to our description of Sufjan Stevens' message is in red. The red box indicates an anomalous case similar to what was discussed in Section 5.4.3.

See for example with preferences μ^*, ν^* presented in Tables 42. In that case, we obtain the following:

$$\mathcal{P}(27a) \approx 0.4446$$

$$\mathcal{P}(27b) \approx 0.1951$$

$$\mathcal{P}(27c) \approx 0.2688$$

In itself this seems to be a good model for the idealized Sufjan Stevens that we have been working with so far. Table 43 presents the interpretations of a cagey listener when hearing one of the messages (with no uncovering of preferences). We can see that upon receiving (27a), the most likely interpretations made by this listener are $\langle w_m, w_{jc}\rangle$ and $\langle \pi_{qs}, \pi_{cs}\rangle$, as is expected from the dual reading we get of Stevens' songs.

CONCLUDING REMARKS ON NON-POLITICAL DOGWHISTLES

The voluntary ambiguity that can be observed in *dogwhistles*, which makes them useful in strategic approaches to discourse, is not limited to the political realm. There are reasons to think that a similar phenomenon is at play in some cultural content. The formal scaffolding we have put forward in Chapter 5 can be re-used to describe these situations, with a slight dif-

$\langle w, w' \rangle$	μ^*
$\langle w_m, w_m \rangle$	1
$\langle w_m, w_{jc} \rangle$	2
$\langle w_{jc}, w_m \rangle$	2
$\langle w_{jc}, w_{jc} \rangle$	1
$\langle \pi, \pi' \rangle$	ν^*
$\langle \pi_{qs}, \pi_{qs} \rangle$	1
$\langle \pi_{qs}, \pi_{cs} \rangle$	2
$\langle \pi_{cs}, \pi_{qs} \rangle$	2
$\langle \pi_{cs}, \pi_{cs} \rangle$	1

Table 42: Possible preference functions for the Sufjan Stevens example. Any case implying a variation in interpretation is preferred. Note that the context where the interpretations are reversed (L_r interpreting w_{jc} and L_c interpreting w_m, for example) are part of the anomalous case. No strongly dispreferred state.

$L_{Cag}(\langle w, w' \rangle \mid m)$	(27a)	(27b)	(27c)
$\langle w_m, w_m \rangle$	≈ 0.2253	≈ 0.5877	0.0779
$\langle w_m, w_{jc} \rangle$	≈ 0.4670	≈ 0.1316	0.1299
$\langle w_{jc}, w_m \rangle$	≈ 0.2198	≈ 0.1754	0.1948
$\langle w_{jc}, w_{jc} \rangle$	≈ 0.0879	≈ 0.1053	0.5974
$L_{Cag}(\langle \pi, \pi' \rangle \mid m)$			
$\langle \pi_{qs}, \pi_{qs} \rangle$	≈ 0.2253	≈ 0.5877	0.0779
$\langle \pi_{qs}, \pi_{cs} \rangle$	≈ 0.4670	≈ 0.1316	0.1299
$\langle \pi_{cs}, \pi_{qs} \rangle$	≈ 0.2198	≈ 0.1754	0.1948
$\langle \pi_{cs}, \pi_{cs} \rangle$	≈ 0.0879	≈ 0.1053	0.5974

Table 43: Results for L_{Cag}. As we did for preference functions, we chose to interpret the anomalous cases from P_S^* as giving out a zero probability on the message that they predict.

ference in the interpretation we make of the objects that are used. Although we have managed to use some flavor of our model to describe the interpretation and production of interpretatively variable messages in cultural content, we should probably not get ahead of ourselves: while it made relative sense to formally describe the political situation, with more easily defined alternative messages, our rendition of the phenomenon when focusing on the cultural content that we've focused on is obviously lacking. On the alternative messages front, strong arbitration was made with regards to both choosing the alternative messages and their interpretations, and a great deal of abstraction/idealization is at play here, at first glance much more so than in the political context explored in the previous chapter. We have also taken a lot of liberties regarding the definition of the semantics of the messages we chose to focus on.

All in all, this approach feels somewhat inappropriate, or at least very incomplete when trying to describe literary production, like we did, and this approach probably does little when it comes to understanding literary processes, either in interpretation or in production.

Although the picture we gave of the interpretation of those literary texts is lacking, the approach does bring us something interesting from a theoretical and formal point of view. One criticism that can easily be made to SMG is that the personae that the model relies on seem very ad hoc, and no process is given for their discovery and/or creation. Our approach to personae is even more ad hoc, yet with the Sufjan Stevens example, we have seen how we can think of duplicitous meaning-making as a way to combine personae into new personae. In the case of Stevens, once we have a cagey listener that interprets, say, $\langle \pi_{qs}, \pi_{cs} \rangle$ from the utterance that they are exposed to, and the two personae are consistent (meaning one is not the strict negation of the other), we can then interpret their combination as being a new persona. In SMG terms, this might be a way of creating personae through the combination of existing personae via the construction of the intersection between the personae's properties into a maximally consistent new set of properties.

6.3 GENERALIZING DWG

In Saul (2019), we are presented with the case of what is called *Protean dogwhistles*. So far we have only explored situations where only two groups were present in the audience. One reason we have done this is that the literature on dogwhistles generally proposes a distinction between ingroup and outgroup members, but this classification is very unsatisfactory, if only because one of the groups is only negatively defined (outgroup members are simply 'anyone who is not part of the ingroup'). It seems much more interesting to divide the audience into more numerous and well-defined groups, groups that in addition could possibly overlap. In the case of dogwhistles like, for example the wonder-working example (3a), the situation that we would have would be that of a dogwhistled message have its implicit interpretation derived by the ingroup (Evangelicals), and not derived by any of the other groups (any group raised in another religious tradition). But why stop there?

The concept of *Protean dogwhistles* as presented in Saul (2019) focuses on the idea that some terms might have a great variety of meanings associated to them, and not merely an ingroup meaning opposed to a general meaning. The idea is already in some form presented in Albertson (2015), where it is noted that religious dogwhistles are in fact interpreted slightly differently by different outgroups, with some perceiving the dogwhistle to some extent. Taking the word "immigration" as it was used during the Brexit campaign in the UK, Saul (2019) takes a further step by acknowledging that there are some words that have an ever-changing reference. In the context of Brexit, when "immigration" was invoked as a threat, it could be used to refer to many groups of migrants of different status and interests. In Saul (2019), the difficulty of countering such terms is underlined, because of their very underdetermined reference. Implementing such Protean dogwhistles in our model would be easy, we could in fact replicate the implementation present in Chapter 5, but making the reference of the dogwhistle term even more vague. We will do so in Section 6.3.1.

Calling these instances *dogwhistles* is however disputable if we refer to the understanding we had of them so far. There is here no *hidden meaning* or *secret handshake*, just highly polysemous words. In fact, it is specifically underlined in Henderson and McCready (to appear) that the model presented therein does not aim at explaining these instances (called there *multivocal appeals*). On our end, we still think that the enterprise might be worth it for the following reasons:

1. The model we've proposed so far is to some extent a disambiguation model, and we think it should be able to handle cases like these.

2. The differentiation between dogwhistles and multivocal appeals is not as clear cut to us than it seems to be to Henderson and McCready (to appear).

Regarding point number 2, the distinction between dogwhistles and multivocal appeals is not done on the basis of the number of possible interpretations available to a given lexical item/sentence. Dogwhistles as we have defined them are not incompatible with the idea that any given dogwhistle might have several *hidden meanings*, which were translated in our model as interpretations only accesible through specific interpretation functions that are themselves not necessarily accesible to all listeners. It is the fact that the meanings of the dogwhistle are *hidden* that makes them dogwhistles. Determining which meanings of a word are hidden, however, is no simple task (even though using a formal model might make it seem so), and the more a word has meanings the more likely it is that at least one of those meanings might be *hidden* to at least some group in the population. Taking up the "immigration" example again, while it certainly has a number of explicit meanings, we might very easily imagine additional hidden meanings whereby, e. g., the word itself would not in fact refer to migrants but to people of foreign origin. Clearly delimiting dogwhistles and mutlivocal appeals is easily done in theory, dogwhistles having one supplementary property of having hidden meaning,

making them a strict subset of multivocal appeals; it is not so easily done in practice, where determining which words exactly have or do not have hidden meaning is an arduous task[13].

In Section 6.3.2, we will define an easy generalization of our DWG over cases with any number of hidden meanings, with an attempt of application to a simple case with 3 different groups.

6.3.1 *Protean dogwhistles*

Let's take as a starting point the following statement from UKIP's manifesto[14]:

(28) Mass immigration is undesirable for two reasons: it has caused cultural division and undue population growth.

Inspired by Saul (2019), we can define the following alternatives:

(29) a. Mass *European* immigration is undesirable for two reasons: it has caused cultural division and undue population growth.

b. Mass *Eastern European* immigration is undesirable for two reasons: it has caused cultural division and undue population growth.

c. Mass *Syrian* immigration is undesirable for two reasons: it has caused cultural division and undue population growth.

d. Mass *Muslim* immigration is undesirable for two reasons: it has caused cultural division and undue population growth.

e. Mass *dark-skinned people* immigration is undesirable for two reasons: it has caused cultural division and undue population growth.

This underlines some of the many meanings ascribed to the notion of "immigration" in Saul (2019). Those interpretations differ widely, with (29a), (29b) and (29c) referencing (sets of) nationalities, (29d) referencing members of a religious group, and (29e) referencing a set of people through their skin-color. If we recall the *norm of racial equality* that we mentioned in Chapter 4, some of those alternatives seem to abide by it by not referencing groups against which there has been prototypical racist oppression:

> *"The Norm of Racial Egalitarianism does seem to hold in the* UK. *And, similarly to the* US *(though not identically) the paradigm case of racism is generally understood to be prejudice of whites avainst darker-skinned people. But [. . .] this is only one of the things that might be dogwhistled by the Leave*

13 Not to mention: the results of such a task would in any case be highly volatile, given that we have little reason to think that there are expressions that cannot and never will become dogwhistles, or that existing dogwhistles might not have all their hidden meanings uncovered.

14 Although the document itself does not date back to the Brexit campaign, the party's position on immigration has not changed. Freely available here: https://irp.cdn-website.com/f6e3b8c6/files/uploaded/Living%20Manifesto%20with%20Updated%20Cover%20REV1.pdf

campaign's invocation of the protean notion immigration. *For other audiences, what is dogwhistled is about other groups [. . .] [a]nd each of these strays importantly from the paradigm case of prejudice against the dark-skinned."*

— Saul (2019)

Taking inspiration from Saul (2019), we can attempt to implement this situation in our model.

Let Γ be a DWG: $\langle \{S, \{L_{1_l}, L_{1_r}\}\}, L_{0_{l/r}}, W, M, \Pi, \textsc{Lex}, \textsc{Soc}, \Pi\text{-}\textsc{Lex}, \Delta\text{-}\textsc{Soc}, Pr_W, Pr_\Pi \rangle$.

Once again, temperature parameters $\alpha, \alpha', \beta, \beta'$ are all set to 1.

- S, the speaker, would in this case be the writers behind the manifesto where (28) is found. For the sake of the example, one can imagine e. g., Nigel Farage saying this.
- L_{1_l}, L_{1_r} are two pragmatics listeners, respectively a *Leave voter* and a *racist Leave voter*
- $L_{0_{l/r}} : M \to \Delta(W), \Delta(\Pi)$ are Literal Listeners corresponding to the priors found below, one for each listener.
- $W = \{w_e, w_{ee}, w_s, w_m, w_d\}$ is a set of possible states of the world, respectively one where the type of immigration referred to by S using (28) is one of the sets of individuals explicitly referenced by any of the (29) alternatives.
- $M = \{(28), (29a), (29b), (29c), (29d), (29e)\}$ is the set of relevant possible messages available to the speaker.
- $\Pi = \{\pi_{rc}, \pi_{nrc}\}$ we re-use the *racist conservative* and *non-racist conservative* personae from Chapter 5, although we're in a different context, they seemed to fit.
- $\textsc{Lex} = \{[\![\cdot]\!]_{rc}, [\![\cdot]\!]_{nrc}\}, \textsc{Soc} = \{[\cdot]_l, [\cdot]_r\}$ we have one interpretation function per persona and one indexation function per listener.
- We set Δ-Soc as follows:

$$\Delta\text{-Soc}([\cdot]_l) = p$$
$$\Delta\text{-Soc}([\cdot]_r) = 1 - p$$

  With $p = 1$ for L_{1_l} and $p = 0$ for L_{1_r}. Again, each listener only has access to one interpretation function.

- We set Π-Lex as follows:

  For L_{1_l}:

$$\Pi\text{-Lex}(\pi_{rc}) = \begin{cases} [\![\cdot]\!]_{rc} : 0 \\ [\![\cdot]\!]_{nrc} : 1 \end{cases}$$

$$\Pi\text{-Lex}(\pi_{nrc}) = \begin{cases} [\![\cdot]\!]_{rc} : 0 \\ [\![\cdot]\!]_{nrc} : 1 \end{cases}$$

For L_{1_r}:

$$\Pi\text{-Lex}(\pi_{rc}) = \begin{cases} [\![\cdot]\!]_{rc} : 1 \\ [\![\cdot]\!]_{nrc} : 0 \end{cases}$$

$$\Pi\text{-Lex}(\pi_{nrc}) = \begin{cases} [\![\cdot]\!]_{rc} : 0 \\ [\![\cdot]\!]_{nrc} : 1 \end{cases}$$

- The priors over worlds Pr_W are uniform. The priors over personae Pr_Π are uniform as well.

Interpretation and indexation functions for this implementation are found in Tables 44 and 45. The L_1 results are presented in Figures 10 for legibility. We can observe a few things, the first being that in this case, the Protean dogwhistle formulation (28) can lead to many an interpretation, as is expected. In terms of content interpretation, we observe that L_r is more likely to lead to a socially marked interpretation like those of (29c), (29d) or (29e) than L_l is, although ultimately the interpretation of (28) leads to absolute uncertainty in terms of content for L_r. A duplicitous speaker having the literal listeners of L_r and L_l as internal listener models would make their utterance choice such that any typical dogwhistle situation, including for example $\langle w_d, w_e \rangle, \langle \pi_{rc}, \pi_{nrc} \rangle$, will lead to (28) to be the utterance that they will most probably use[15]. The results for $\mathcal{P}(m)$ in that case are as follows:

15 Unfortunately, due to the very large amount of contexts available to the duplicitous speaker here, there is no practical way to vizualize it, may it be in table or figure form. this is in part due to the fact that the separation of propositional meaning into this many possible worlds leads to an explosion in possible combinations. Appendix B however presents a way for anyone to reproduce all these computations.

$[\![m]\!]$	$[\![\cdot]\!]_{rc}$	$[\![\cdot]\!]_{nrc}$
(28)	$\{w_e, w_{ee}, w_s, w_m, w_d\}$	$\{w_e, w_{ee}, w_s\}$
(29a)	$\{w_e, w_{ee}\}$	$\{w_e, w_{ee}\}$
(29b)	$\{w_{ee}\}$	$\{w_{ee}\}$
(29c)	$\{w_s\}$	$\{w_s\}$
(29d)	$\{w_m\}$	$\{w_m\}$
(29e)	$\{w_d\}$	$\{w_d\}$

Table 44: Definition of the interpretation functions for the Protean dogwhistle example.

$[m]$	$[\cdot]t_r$	$[\cdot]_l$
(28)	$\{\pi_{rc}\}$	$\{\pi_{nrc}\}$
(29a)	$\{\pi_{nrc}\}$	$\{\pi_{nrc}\}$
(29b)	$\{\pi_{nrc}\}$	$\{\pi_{nrc}\}$
(29c)	$\{\pi_{rc}\}$	$\{\pi_{rc}\}$
(29d)	$\{\pi_{rc}\}$	$\{\pi_{rc}\}$
(29e)	$\{\pi_{rc}\}$	$\{\pi_{rc}\}$

Table 45: Definition of the indexation functions for the Protean dogwhistle example.

$$\mathcal{P}(28) \approx 0.2439$$

$$\mathcal{P}(29a) \approx 0.2389$$

$$\mathcal{P}(29b) \approx 0.2310$$

$$\mathcal{P}(29c) \approx 0.0951$$

$$\mathcal{P}(29d) \approx 0.0956$$

$$\mathcal{P}(29e) \approx 0.0956$$

These tendencies can again be modulated by the use of preference functions.

6.3.2 *Generalized DWG*

The Protean dogwhistle case is interesting in the sense that even though the uncertainty is a lot more important on the listener side, the duplicitous speaker, when in a context adequate for dogwhistling, will use the most dogwhistley alternative available, almost invariably. The reason behind this is that the dogwhistle term itself is highly polysemous including to the ingroup. This does not sound out of place in this precise example, but we can nonetheless imagine a situation where that polysemy is constrained by envisioning less monolithic in-group/outgroup distributions, in fact rendering the distinction itself caduceous. If we take up

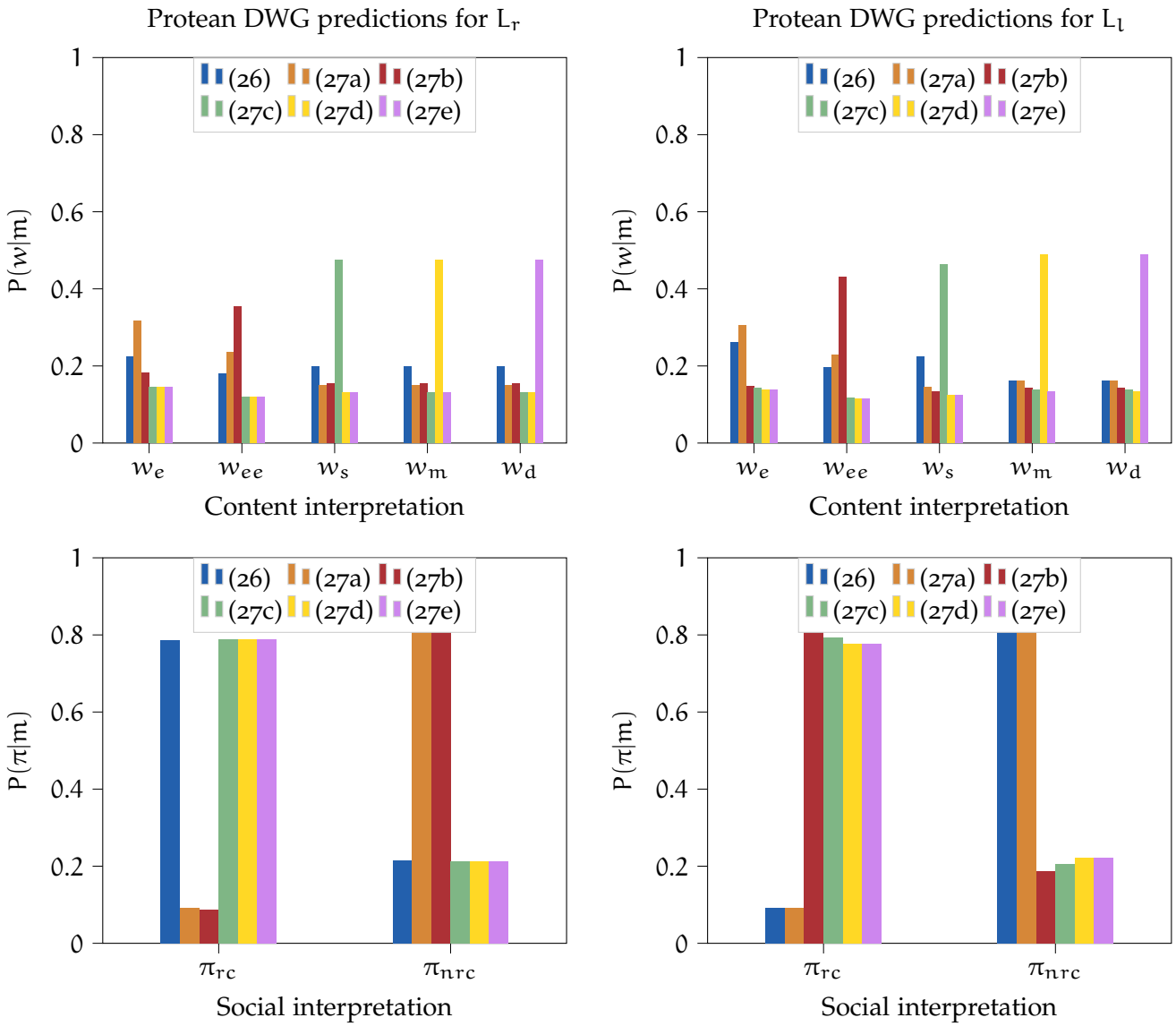


Figure 10: Visualization of pragmatic listeners for the Protean dogwhistle case. on the left is L_r, and on the right is L_l. Top row is content interpretation, bottom row is social meaning interpretation.

the example of (28), we could envision that according to their group, listeners only assign one of the many meanings of the expression. We would end up with, say, a group that identifies as *white supremacists* and see the individuals referenced to when mentioning "immigration" in a derogatory way as being people with darker skin; another group that identifies them with people from outside Europe; yet another group that identifies them with Eastern Europeans, etc.

The model as we have defined it in Chapter 5 only takes into account two possible players, but we can easily modify it to account for any number of players, and so can solution concepts be appropriately modified to do so.

Definition 6.3.1. Generalized Dogwhistle game

A Generalized Dogwhistle game is a tuple of the form:

$$\langle \{S, \mathbb{L}, L_0, W, M, \Pi, \textsc{Lex}, \textsc{Soc}, \mathfrak{P}\text{-}\textsc{Lex}, \Delta\text{-}\textsc{Soc}, Pr_W, Pr_\Pi \rangle$$

In this tuple, we find the following:

1. S is the speaker.
2. $\mathbb{L}$ is a set of listeners $\{L_{k_0}, \ldots, L_{k_n}\}$ of arbitrary finite size such that $|\mathbb{L}| \geqslant 2$.
3. L_0 is the Literal Listener, an application $M \to \Delta(W), \Delta(\Pi)$.
4. W is a set of worlds w.
5. M is a set of messages m.
6. Π is a set of personae π.
7. $\textsc{Lex}$ is a set of *lexica*, or interpretation functions $[\![\cdot]\!]$.
8. $\textsc{Soc}$ is a set of *social lexica*, or indexation relations $[\cdot]$, as per Burnett (2017).
9. $\mathfrak{P}$-$\textsc{Lex}$ is a set of functions Π-$\textsc{Lex}$: $\Pi \mapsto \Delta\textsc{Lex}$ which map specific personae π to probability distributions over available lexica in $\textsc{Lex}$. There is one such function per listener.
10. Δ-$\textsc{Soc}$ is the set of priors over indexation relations $[\cdot] \in \textsc{Soc}$ there is one such prior distribution per listener.
11. Pr_W is a set of probability distributions over W, representing listener prior beliefs regarding the state of the world.
12. Pr_Π is a set probability distributions over Π, representing listener prior beliefs regarding the persona of the speaker.

$[\![m]\!]$	$[\![\cdot]\!]_{nrc}$	$[\![\cdot]\!]_{rcw}$	$[\![\cdot]\!]_{rci}$
(28)	$\{w_e, w_{ee}, w_s\}$	$\{w_d\}$	$\{w_m\}$
(29a)	$\{w_e, w_{ee}\}$	$\{w_e, w_{ee}\}$	$\{w_e, w_{ee}\}$
(29b)	$\{w_{ee}\}$	$\{w_{ee}\}$	$\{w_e, w_{ee}\}$
(29c)	$\{w_s\}$	$\{w_s\}$	$\{w_e, w_{ee}\}$
(29d)	$\{w_m\}$	$\{w_m\}$	$\{w_e, w_{ee}\}$
(29e)	$\{w_d\}$	$\{w_d\}$	$\{w_e, w_{ee}\}$

Table 46: Definition of the interpretation functions for the Generalized DWG example.

$[m]$	$[\cdot]t_l$	$[\cdot]_w$	$[\cdot]_i$
(28)	$\{\pi_{nrc}\}$	$\{\pi_{rc}\}$	$\{\pi_{rc}\}$
(29a)	$\{\pi_{nrc}\}$	$\{\pi_{nrc}\}$	$\{\pi_{nrc}\}$
(29b)	$\{\pi_{nrc}\}$	$\{\pi_{nrc}\}$	$\{\pi_{nrc}\}$
(29c)	$\{\pi_{rc}\}$	$\{\pi_{rc}\}$	$\{\pi_{nrc}\}$
(29d)	$\{\pi_{rc}\}$	$\{\pi_{rc}\}$	$\{\pi_{nrc}\}$
(29e)	$\{\pi_{rc}\}$	$\{\pi_{rc}\}$	$\{\pi_{nrc}\}$

Table 47: Definition of the indexation functions for the Generalized DWG example.

Due to the way the model is defined, adding more listeners quickly leads to an explosion of the contexts under consideration for e. g., a S_{Dup} solution concept, the utility functions U_S^* and choice rule P_S^* need to be rewritten as follows:

$$U_S^* \mapsto U_S^*(\langle \pi_0, \ldots, \pi_n \rangle, m) = \sum_{L_{k_n} \in \mathbb{L}} \log(P_{k_n}(\pi|m)) - C(m)$$

$$U_S^* \mapsto U_S^*(\langle w_0, \ldots, w_n \rangle, m) = \sum_{L_{k_n} \in \mathbb{L}} \log(P_{k_n}(w|m)) - C(m)$$

$$P_S^*(m|\langle, w, w'\rangle, \langle \pi, \pi' \rangle) \mapsto P_S^*(m|\langle w_0, \ldots, w_n \rangle, \langle \pi_0, \ldots, \pi_n \rangle)$$

Let's try to use this version on the Protean dogwhistle case, but this time involving 3 listeners: $\mathbb{L} = \{L_l, L_w, L_i\}$, respectively a *standard leave voter* a *white supremacist leave voter* and an *islamophobic leave voter*[16]. L_w and L_i come with their version of the $[\![\cdot]\!]_{rc}$ interpretation function, and everyone gets their own indexation function, presented in Tables 46 and 47. The rest is otherwise identical to what we had in Section 6.3.1. The results for pragmatic listeners are found in Figure 11. As expected, we find that the interpretations by the listeners of an utterance like (28) becomes more specified among the two ingroups.

16 This is for the sake of the argument, it goes without saying that those last two categories are hugely permeable to each other.

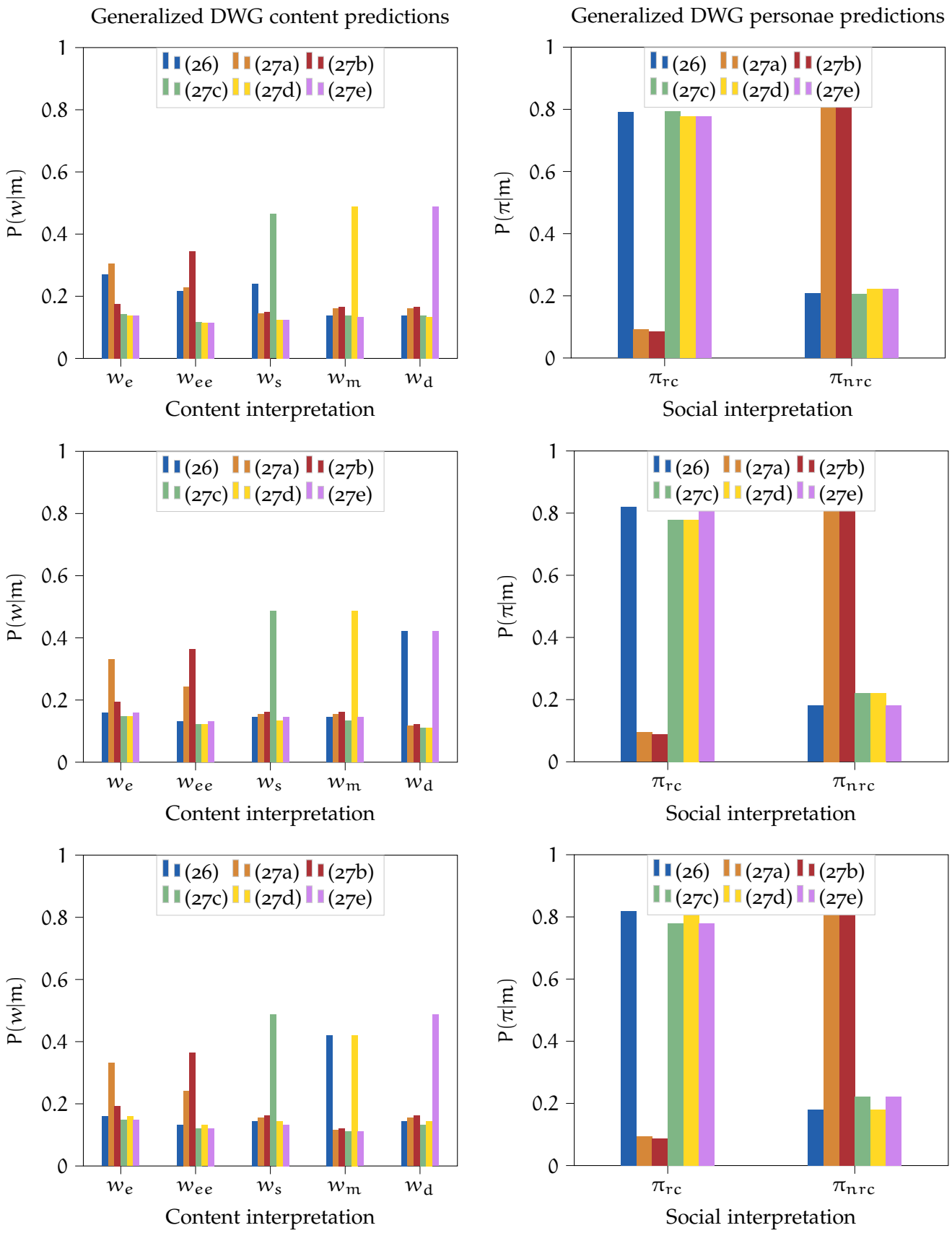


Figure 11: Visualization of pragmatic listeners for the Generalized DWG case. From top to bottom, we have L_1, L_w, L_i. Content interpretations are on the left, personae interpretations on the right.

6.4 CONCLUDING REMARKS

In this chapter, we have attempted to generalize the DWG model to other situations, that were either qualitatively different, but sometimes referred to as "dogwhistles" (Sections 6.1 and 6.2), or different in a sense of scale, though ultimately closer in spirit to traditional dogwhistles (Sections 6.3.1 and 6.3.2), leading us to a more general model in Section 6.3.2. We have seen, for the first case that a different interpretation of the various mathematical objects we have created could lead some understanding of situations that are intuitively very different from dogwhistles. For the second case, we have seen that the model can handle situations that are more complex, in terms of having more ideal listeners as well as in terms of having more polysemous terms, with very few modifications.

This has allowed us to take advantage of one of the main contributions of formal modeling: bringing together seemingly very dissimilar phenomena and treat them in what are essentially identical ways, allowing us to reassess initial judgments about their (dis)similarity.

The question begs to be asked, however: is this model a good model for dogwhistle communication? It does make a number of predictions correctly both as a listener and as a speaker model, but we've seen already in Chapter 5 that it is severely limited. The model overall is an attempt to translate two different models (RSA and SMG) into one and have them interact. This by itself comes with a lot of problems. What we have in reality here is a speaker that attempts to play two games with a single move, and listeners that are aware of this. Unless we add suppressing parameters like we did in Section 5.4.1, then the model will not converge on a proper equilibrium, no matter how many times we repeat the recursive solution concept of standard RSA. In this chapter, we've also seen that, being based on RSA, this model revolves around exploiting the semantics of competing utterances to enrich the semantics of underdetermined utterances, but this begs the question of how to determine which utterances can be considered relevant competing utterances. In the case of, say, scalar implicatures, the question is more easily answered; in the case of SMG applied to phonetic variants, then there are inventories of how one specific utterance might be uttered, but in our case, the actual number of competing utterances is potentially limitless. This leads to cases like those in Sections 6.1 and 6.2 to be treated in a rather unsatisfactory way, or leading to a multiplication of anomalous cases (Section 6.3.2).

Before going into the rest of the dissertation, some more in-depth remarks about this type of formal modelling in general can be made. In the last two chapters, we have put forward a formal model for the description of dogwhistle communication. While formal models can bring new insights and a new understanding of the phenomena they describe by likening them to other, more easily understood phenomena, those models are sometimes hard to interpret. How should we understand the predictions of the DWG model? What does the *probability of interpreting* such content from such message even mean?

There are several understandings of the concept of "probability" in mathematics and philosophy[17], but one that is very generally used, particularly in experimental sciences, is the *frequentist* interpretation of the concept, whereby the probability of an event occurring is the frequency of that event effectively occurring over many trials. Such an interpretation of the concept was put forward, notably, by mathematicians like Venn, Pearson or Fisher.

More precisely, it means that the probability of a situation s producing an outcome o is the limit of its relative frequency over an infinite amount of trials. This interpretation of the concept allows to make sense easily of, say, the odds of throwing an unbiased six-sided die and landing on a 6. Over just a few trials, it can be that one lands on 6 many times, but as the number of trials increases, the observations will be such that the number of times one lands on a 6 will gradually approach $1/6^{th}$ of the total number of trials. If we went on indefinitely, it would theoretically be the case that exactly $1/6^{th}$ of the trials have had 6 as an outcome. This makes a lot of sense for situations like the throwing of dice, because it is easy to imagine a replication of the experiment with minimal alterations.

A frequentist interpretation of the results of our model would essentially be as follows:

LISTENER MODEL For the listener model, a frequentist interpretation of the output of the model would be that over a great number N of occurrences of an utterance m directed to a listener, interpretations w and π would be respectively triggered $P(w|m) * N$ and $P(\pi|m) * N$ times. Similarly, we can see it as being just a single utterance directed at very many listeners, with the relative frequencies of interpretation in the crowd being the same.

SPEAKER MODEL For the speaker model, a frequentist interpretation of the output of the model would be first that given information w, π that the speaker needs to convey, they will use a message m $P_S(m|w, \pi) * N$ times, with N being the number of times that they need to convey the context w, π. Another interpretation we can make of the output is that when confronted to any of the possible contexts, with all contexts being equally likely, a speaker will use m $\mathcal{P}(m) * N$ times, with N the number of situations where an utterance is called for.

The main advantage of a frequentist reading of the outputs of our model is that it makes it very easy to transform those predictions into *quantitative* predictions and then test them experimentally or through corpus analysis. This is notably what is done for SMG in Burnett (2019), using the number of occurrences of the variant [in] over [iŋ] according to contexts in Barack Obama's utterances. The parameters of the model are then tuned to make it fit the observed data, and this tuned model can be used for further predictions. This also happens to be the general understanding of how a *mixed equilibrium*[18] would work over repeated interactions.

17 Note that the branch of mathematics, probability theory, does not bother itself with these philosophical interpretations as much as one would think. Specifically, probability theory is focused on the valid operations that one can do on probability distributions and the theorems that follow, but it has no particular interest in how to assign probabilities to events in the first place, which is closer to what is understood by *interpretation of probabilities*. This question is closer to a philosophical question than to a purely mathematical question.

18 See Appendix A.

But in our case (as is the case for both SMG and RSA), there are limitations to such an approach. One of those limitations lies in the fact that the notion of "reproducibility" feels inappropriate. Unless one considers extremely abstract cases, there is no such thing as replicating an utterance, an utterance is characterized not simply by the linguistic form that is used during it, but also by its context of enunciation. Regarding RSA, this interpretation also leaves us unclear with regard to what the model predicts people will interpret. The solution concept from RSA is recursively defined, and one key result from this framework is that the implicature mechanism emerges after very few recursive steps, but also that the model eventually converges on the pragmatically enriched interpretation. One thing that is left unclear is: how is the temperature parameter α defined; after how many recursive steps do we consider the model to be a good representation of cognitive processes[19]. When we decide to stop (say, at L_1), how are we to interpret that result and how do we tune the model to account for existing data? One key thing about generalized conversational implicatures like scalar implicatures is that they are said to always take place when in the right context[20], so surely a frequentist interpretation of those results would be unsatisfactory, unless the odds of interpreting *not all* when hearing "some" are extremely high (which only happens with very high α or after several steps of recursive reasoning).

Regarding SMG, the frequentist interpretation allows us to account for existing data in Burnett (2019), but the nature of what is being interpreted is unclear, the personae and their traits are mostly defined *ad hoc*, and in that case it is not so much the probabilistic nature of the interpretation that calls for questioning as it is the content of the interpretation itself, since the model can only make predictions among the set of possibilities laid out by the scientist behind it, no matter how appropriate they happen to be. In case we abide by the frequentist reading, the result is therefore as obscure as before, at least regarding the listener model, and there is great uncertainty on the message that is effectively conveyed.

Needless to say that a model like DWG that combines both RSA and SMG will present the same issues. In addition to all of this, we have chosen to specifically interpret the output from the model in a *non-frequentist* way. Probability scores as they appear in the model have so far only been interpreted as listener/speaker *uncertainty*, and not as an actual frequency of occurrences, in a way that might remind one of the so-called *Bayesian* interpretation of probabilities.

Philosophically that makes DWG very different from the other approaches, and practically, choosing to interpret the output in this way forbids us to make the kind of quantitative predictions that are warranted by frequentist approaches. Testing the results of the model with this interpretation of probabilities implies that listeners do not simply choose an interpretation among possible interpretations, but rather assign a degree of belief to each possible interpre-

19 In the case of what we have seen, e. g., in Section 2.3.2, these two questions are equivalent, a higher α will make the model converge faster/in fewer steps towards the pragmatically enriched meaning.

20 At least after a certain age, see e. g.Noveck (2001).

tation. In other words, the philosophical position that is defended by this interpretation of the model is that the meaning of utterances *is probabilistic in nature*.

We already said something similar about the way RSA transforms deterministic semantic truth conditions into probabilistic pragmatic interpretation, but having an interpretation of the output of the model in a view closer to Bayesianism than to frequentism is in fact the way to go if one wants to treat *meaning* as probabilistic.

Like many formal models, DWG calls for an extreme abstraction of the real-world situations that we attempt to describe, sometimes to the point of completely denaturing them, and is in this respect not entirely satisfactory. It does shed light however on how one can attempt to generalize models that are initially conceived for the description of very specific phenomena. Trying to construct formal models of phenomena can also serve as a means to see which specific concepts are useful for the description of the phenomena in question, even when the model itself makes uncanny predictions.

In the case of dogwhistles, DWG is a model that insists on thinking of them in terms of dialectal variation and *synonymy*. More specifically, in the next chapter, we will talk of *asymmetric synonymy*, what that can mean for dogwhistles, and how that notion can be approached using different methods.

Part IV

DOGWHISTLE DETECTION AND COMPUTATIONAL DEFINITION

This part focuses on tentative computational approaches to the description of the dogwhistle phenomenon. Notably, we try to find ways to automatically detect dogwhistles in political corpora. We also zoom out from the formalist perspective to discuss the concept of *dogwhistles* and see in which ways they do or do not correspond to the formal definitions we have put forward.

7

DOGWHISTLES, SYONYMY AND DISTRIBUTIONAL SEMANTICS

"The placing *of a text as a constituent in a context of situation contributes to the statement of meaning since situations are set up to recognize* use. *As Wittgenstein says, 'the meaning of words lies in their use.' The day to day practice of playing language games recognizes customs and rules. It follows that a text in such established usage may contain sentences such as 'Don't be such an ass!', 'You silly ass!', 'What an ass he is!' In these examples, the word* ass *is in familiar and habitual company, commonly collocated with* you silly–, he is a silly–, don't be such an–. *You shall know a word by the company it keeps!"*

— The famous sentence by John Rupert Firth. In context. (Firth, 1957)

"'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe."

— *Jabberwocky*, a poem by Lewis Carroll. Although most content words in the poem are replaced with gibberish, the meaning of some still seems to transpire. "Somehow it seems to fill my head with ideas – only I don't exactly know what they are!" (Carroll, 1872)

The work presented in the next two chapters is based on and expands upon Dénigot and Burnett (2021)

The formal model that we put forward in the last two chapters is useful to give us an insight into how one might describe dogwhistles in the sense of "what linguistic concepts and mathematical objects can be usefully invoked when attempting a description of dogwhistles", but as underlined in Section 6.4, such formal descriptions can quickly become hard to interpret, and while game theoretic pragmatic models can give us predictions on the production and interpretation of dogwhistles, which is all we need for simpler cases like scalar implicatures, they do not give us a way to *identify* them. Contrary to instances of cooperative communication, dogwhistles rely on duplicity, and as such identifying them is a task in itself. That task is irrelevant for the other items which we've viewed under the light of game theory: their presence is either easily attested (scalar implicatures) or assumed (socially meaningful statements). The formal model we have put forward describes situations that could be characterized as *dogwhistle communication*, but it does not in fact give us a reliable way to identify whether a specific instance of, e. g., a word can be construed as being a dogwhistle.

In this chapter, we will explore ways to give a definition of dogwhistles that could then potentially be used for the (automatic) detection of dogwhistles, or at least give a hint towards an answer to the question of identifying them.

7.1 COMPUTATIONAL DEFINITION OF DOGWHISTLES: WHY, HOW?

Before attempting to give a definition of dogwhistles that can easily be translated into an (algorithmic) methodology for their detection, we should question the reasons why we would want to do such a thing, and whether it even is possible. We will refer to both questions as the WHY and the HOW.

WHY There are several reasons why we would want to have an algorithmic way to identify dogwhistles in a text or a speech. One of those reasons would basically be linked to conviction in a democratic political ideal: we have seen in Chapter 4 that a number of authors (Goodin and Saward, 2005; Stanley, 2015) see dogwhistles as a direct threat to a healthy democracy, either because they weaken mandates, by potentially making it unclear to voters which policies dogwhistling candidates actually support, or for example because they usher into power figures with anti-democratic views under democratic pretense. In any case, there are political reasons to want to detect dogwhistles (and subsequently find adequate stategies to fight them)[1].

Beyond reasons like these, there are more academic-centered reasons to want to do such a thing. The examples that we have presented in Chapters 4 and 5 have been discussed a lot in the literature on dogwhistles. They present certain properties that make them good examples of dogwhistles, but their main quality with regards to the scholarship on dogwhistles is the fact that they have been properly identified as such. Finding whether a word or a formulation is dogwhistley is an understandably difficult task, which forces one to rely on a deep knowledge about political discourse, political ideologies, and due to plausible deniability is ultimately a judgment call. For the study of dogwhistles itself, it might be interesting to find out whether some methods could lead to evidence regarding the possible dogwhistleness of, e. g., a word, or sentence, or hashtag, etc. While it is likely that no automatic/quantitative approach will lead to results that are more convincing or reliable than an utterance-specific qualitative approach in this specific task, maybe there is still a way to use automatic methods and statistical measures to isolate some potential candidates for dogwhistling, bring in evidence that does not directly rely on personal judgment, and therefore ease the task of detection and further study.

We are not the first to propose something like this. Focusing on Critical Discourse Analysis (CDA), for example, it typically involves "(a) finding a regular pattern in a particular text or set of texts [...] and then (b) proposing an interpretation of the pattern, an account of its meaning and ideological significance" (Cameron, 2001, p. 137). The step of identifying such regular patterns in discourses is usually done manually (see for discourse on LGBT+

1 One argument could also be made that dogwhistles in one form or another are themselves a part of political discourse, and that there is no way to "fight" them *per se* without fighting the very idea of political discourse. To which one could answer that in that case, there is still some interest to be found in detecting and fighting some dogwhistles, specifically those that do not align with a given political ideology.

communities, e. g., Provencher, 2011; Van der Bom et al., 2015; VanderStouwe, 2013). However, many researchers have found computational methods to be more systematic (Mautner, 2016). We argue Word Embeddings (WE) can be used to identify discourses that previous methods miss[2].

Finally, there is a philosophical argument to be made for this question that we have already made in Chapter 1 and is best summed up by a quote we already called upon then:

> *"If we want to understand nature [. . .] then we must use all ideas, all methods, and not just a small selection of them."*
>
> — (Feyerabend, 1975)

Computational tools are merely a set of tools among others, and using them can help in having an all-around understanding of dogwhistles. Specifically, they can possibly bring us something that philosophical (Chapter 4) or formal (Chapter 5, Chapter 6) approaches cannot, although possibly at the expense of some dimensions of understanding. Focusing this time on the detection task might be one way to discover yet other dimensions to dogwhistles and their use, but might also be a way to test the limitations of standard computational tools and computational reasoning. Note that the task of dogwhistle detection is intuitively closer to *discourse analysis* (and in fact to qualitative discourse analysis) than it is to any form of computational approach. The goal here, and this will be repeated several times, is *not* to propose a substitute for careful qualitative discourse analysis, but to propose an alternative view of the problem. Whether this view is useful (or even successful) will be discussed, but it is in no way to be seen as a replacement for existing methods.

HOW The question of whether the task is in itself not a silly endeavor/unattainable goal is not easy to answer. On the one hand, it seems implausible to even think about automatically detecting dogwhistles: we have seen in Chapter 4 that a large part of how they are defined is in their perlocutionary effects, which go beyond any form of linguistic data that we could possibly gather. In Chapter 5, some emphasis was put on the *preferences* and *intentions* of a speaker using dogwhistles (or at least the intentions and preferences that can be retrieved by the listener), which also constitutes data that is unattainable or impossible to gather. What we would need is a characteristic of dogwhistles that can somehow be found using observable linguistic data, meaning that there would be some property of dogwhistles that is due to the signal itself and not just found in the intention or reception of said signal.

The model presented in Chapter 5 uses the idea that dogwhistles change meaning according to the lexicon that is used by the speaker/listener pair. In one lexicon, the dogwhistle term can mean one thing, and in another it can mean another thing, while non-dogwhistle terms would have a more stable meaning across lexica. In other words, in one lexicon, the dogwhistle

2 This point will be illustrated in Chapter 8, using the 2013 debates on *le mariage pour tous* at the French Assemblée Nationale.

is closer to a given meaning, and in another lexicon it is closer to another meaning. This comparison to alternatives is not necessarily doable with the signal taken by itself, but those alternatives might appear in other signals by the same or a similar speaker. If we had access to a large number of signals and had a way to compare the meaning of words that we suspect of being dogwhistles with the meaning of their possible alternatives, then this might constitute a first step towards an attempt to identify dogwhistles in a text. Although the information we are looking for is not necessarily in the signal *per se*, it is in the overall set of signals one has had access to over time.

What we would need to do this is a computationally meaningful definition for the *meaning* of a lexical item and a way to compare that meaning with the meaning of other lexical items or the same lexical item used in other contexts. One possible answer to the first part of this requirement can be looked for in the notions of Distributional Semantics (DS), Vector Space Model (VSM) and Word Embeddings (WE). These concepts are presented in a bit more detail in Appendix C. The way we intend to use these concepts will be explored in Section 7.3. In a few words: using tools from Natural Language Processing (NLP) and machine learning techniques based on ideas usually associated with, notably, Harris (1954), there are ways of assigning series of numbers to words in a meaningful manner, such that if one interprets these numbers in terms of coordinates in a space of many dimensions, then words that are related from a meaning point of view will tend to appear in the same regions of that space. Appendix C presents all of these ideas in more detail; the work presented in this chapter and the next assumes some degree of familiarity with these concepts.

Regarding the issue of comparison, it will be discussed in Section 7.3.3 and then further discussed in Chapter 8, where a solution using the similarity measures provided by a VSM is proposed.

7.1.1 *Some limitations of the computational approach*

Before making serious attempts at this task, there are a number of limitations to the methods we are about to discuss that should be tackled. The main issue is that of *data*, both from a qualitative and a quantitative point of view. Regarding the qualitative point of view, we would need to have a dataset that consists in discourse where the speakers are easily identifiable and where some inference can be made with regards to their ideological group. We think to have found such a dataset in French Parliamentary debates, this will be discussed further in Section 8.1.

Regarding quantitative limitations, the computational tools that we are about to use typically rely on enormous amounts of data, and the amount of data that is used in building the models resulting from these methods is generally strongly correlated to the quality of the model, 'quality' being here broadly defined as the model behaving in expected ways and presenting as few oddities as possible. In our attempt, we do not possess nearly as much data

to work on as these models usually require. While there are ways to circumvent this issue to some extent, we have to expect that our results might not be as impressive as what we would initially naively expect. This is one key reason for us to state that the method we are proposing is not to be understood as a substitute for careful qualitative analysis of the phenomenon we are interested in, we merely envision this as a supplementary way to explore the data that we are given.

Although we do not have access to as much data as one would need to properly train our models due to the very specific nature of the task at hand, the nature of the methods still allows us to consider working on more data than usual discourse analysis tasks. Considerations about dataset size and what they entail will be taken up in Section 7.3.2 and Section 8.2.

Another key limitation of the approach is, again, the difficulty of properly interpreting the outputs from our models. The task we are about to present here is very different from the standard notion of *task* in NLP, in the sense that even though it could resemble a *classification* task, meaning that given a word, the algorithm would classify it as either being a dogwhistle or not being a dogwhistle, this is not how we chose to tackle this, and for several reasons:

1. As stated earlier, our goal here is not to give a method that would replace proper qualitative analysis, but to give a method that would give hints with regards to the status of the lexical units one might be interested in. This is not incompatible with something like a *dogwhistleness score*, which is closer to what we did, but making sure that no threshold was chosen to determine what was or was not a dogwhistle, as this is ultimately not a choice to be made by a machine. What we wanted to do was focus on a different understanding of WE and similar NLP tools, try to use them for *data exploration* rather than applying them to proper tasks.

2. Beyond this philosophical statement, another reason why we are *not* proposing a classification task is that through this computational exercise, we have realized that in fact many things needed to be properly addressed before this could be meaningfully turned into a task with a measurable notion of success. These issues will be addressed mostly in Chapter 8, but will already be tackled in Section 7.3.3.

3. Ultimately, we are not as interested in the accuracy of a potential classifier for dogwhistles as we are in finding meaningful ways to use available data to understand more about the phenomenon of dogwhistle communication. Specifically in that case, WE appear to be a philosophically motivated tool for the study of dogwhistles mostly because they provide us with a well-defined notion of *semantic similarity*.

So the endeavour can be seen as two-fold: first, we wanted to see whether seeing the data through the lens of WE could help us in the detection and description of dogwhistles; second, we wanted to see whether using WE in a way decidedly different from their typical uses could lead to more applications of this technology to questions in linguistics and discourse analysis.

In both cases, we will see that the results do not necessarily point towards the direction of a definite use for WE when tackling those questions, although the process itself might prove fruitful for other similar endeavours. In the case of this work, we have used the *word2vec* algorithm for the generation of WE (Mikolov et al., 2013). The way it works is very briefly explained in Section C.3.

7.2 DOGWHISTLES AS ASYMMETRICAL SYNONYMS

Following what was said in Section 6.4, we will argue here that the concept of *asymmetric syonymy* can be a useful concept to invoke when discussing dogwhistles, especially in the approach to the concept that underlies the formal model from Chapter 5 and Chapter 6. But what exactly do we mean by this?

The principle behind this is best illustrated by looking at Table 16, which we used to define the interpretation functions for the our messages in the "inner cities" example from Section 5.4. Looking at the interpretation functions, we see that the way we have described them, the dogwhistle message m_{dw} is semantically synonymous to either m_{aa} or m_{cc} according to which interpretation function we choose to focus on, meaning it has the same truth-conditional meaning as either one of those. Dogwhistles constitute *synonyms* in the sense that they share their meaning with other utterances, but the synonymy they trigger is *asymmetrical* in the sense that they are not synonymous with the same words across communities.

This concept is in keeping with how we have treated dogwhistles so far as well as with the approaches presented in Chapter 4. Remember that dogwhistles are initially thought of as:

> *"[A] way of sending a message to certain potential supporters in such a way as to make it inaudible to others whom it might alienate or deniable for still others who would find any explicit appeal along those lines offensive."*
>
> — (Goodin and Saward, 2005)

This is ordinarily understood as dogwhistles being a way to send *two* messages, an explicit, accessible message and an implicit unaccessible and less socially acceptable message. We have already used several times the notion of a *verbal secret handshake*, which even though it is a useful paraphrasing for what is happening insists a lot on the idea that there would be a notion of *general discourse* where the dogwhistle would have no particular signification as opposed to a *secretive discourse* where it does. What we want to insist upon here, like we insisted on in Section 6.3.2 is that this *secretive* meaning is not necessarily limited to one community, and that in fact dogwhistles have one possible meaning in one community, another one in another, yet another one in yet another, etc.

This can easily be translated in our formal framework, and we have done something in that vein in Section 6.3.2. Analyzing dogwhistles under the light of dialectal variation as we did automatically leads to this. Dogwhistles under our understanding would be words (or

sentences, expressions, etc.) that are judged to be semantically similar to different other units according to one's preferred dialect.

This understanding allows us to do two things: discuss dogwhistles in terms relating to WE, which is what we will do in Section 7.3; assign dogwhistleness to something *measurable*, where all words would have a certain *dogwhistleness* score[3], and 'more dogwhistley' words would score higher (for example, in virtue of showing more variability in their semantic neighbors across communities).

The next sections will be devoted to giving a more concrete rendition of these ideas in WE terms as well as giving the underlying reasons why we think WE are philosophically motivated in treating this issue. Knowing the basics of WE will be necessary for a proper understanding of this, see Appendix C for a short introduction.

7.3 WORD EMBEDDINGS AS A COMPUTATIONAL APPROACH TO SYNONYMY

For the last few years, word embeddings have been used for a variety of tasks, from document classification (Kusner et al., 2015) to sentiment analysis (Yu et al., 2017). Using the philosophical insights of Distributional Semantics (DS), they allow us to give an intuitive mathematical representation of words and their relationships to each other. They notably allow us to quantify the meaning differences of two words in their vector space using basic similarity metrics (cosine similarity), which has come with a number of interesting properties, including seemingly capturing non-trivial relationships between words, such as gender in some cases[4].

This is an important contribution of DS from a theoretical perspective : it brings with it a computational definition of word similarity. Once a VSM is constructed, all the word vectors in it are separated from one another by a given distance. Given the way the space is structured, the geometric distance between words conceptually corresponds to semantic similarity, with words with references that are intuitively conceptually close being closer to one another in the space. In a pure and abstract understanding of this, pairs of words that perfectly align in the space (with a cosine similarity of 1) would be understood as appearing in the exact same contexts, all the time, and therefore be considered synoyms in the distributional sense.

Such a situation probably does not exist, it is very hard to even find examples of pairs of words that are perfect synonyms. The notion of "synonymy" is in practice made more or less redundant from an empirical standpoint with a VSM, and this is a good thing, it transforms an absolute relation that has very little relevance in the real world but has an intuitively clear meaning into something *continuous*: seeing distance as semantic similarity, there are no such things as *synonyms*[5] in DS, there are just words that are more or less close to each other in the space. All words are similar *to some degree*.

3 There probably aren't many words whose semantic neighbors do not differ across communities.
4 See the now famous example in Mikolov et al. (2013) where *king - man + woman = queen*.
5 Or antonyms for that matter.

7.3.1 *Bias in word embeddings*

Importantly, a VSM is built using a corpus of actual utterances[6], the meanings of the words that we derive from the space are contingent to the choice of that corpus, and the production and use of words changes over time. Without surprise, this has an effect on the embeddings if we pay attention to corpora of different eras. Measures of similarity have been used to assess semantic variation of words through time using large corpora in a number of works (Garg et al., 2018; Hamilton, Leskovec, and Jurafsky, 2016; Rudolph and Blei, 2018), seeing how certain abstract concepts have historically been associated with different groups, concepts or issues across time. While this property has been used successfully in some research, it is a very important problem in many applications. Recently, there has been much talk of the issue of *bias* in AI systems (Bolukbasi et al., 2016; Garg et al., 2018; Gonen and Goldberg, 2019; Manzini et al., 2019; Zhao et al., 2019), whereby a corpus exhibiting a certain ideology that can easily be embedded in the relations between the word vectors will in fact lead to undesirable, ideologically-laden similarity patterns.

This talk of *ideologically-laden* language deserves some more discussion. What does this actually mean? In practice, works on these issues focus on a *racist* or *sexist* bias in the embeddings whereby, e. g., some jobs will be more readily associated with women than men, or more readily associated with black people over white people, in ways that reflect the presence of sexist and racist oppressions in society. It is important to underline however, that there is no such thing as *ideologically neutral* language. The issue here is not that the embeddings display ideological associations, the issue is that these specific associations constitute a problem. From a political and philosophical standpoint, it is important to state this since a lot of the research around the issue has focused on "reducing bias" or "debiasing" embeddings, but there is no such thing as this, there is no solution to bias, bias is a property of the VSM itself. Adding more and more data will not solve bias, in fact, in a society where oppressions are *systemic*, adding more data will just result in more of the same.

This leads to another point: the issue of bias is usually linked to smaller corpora, but it is in no way specific to them, it is present in all subfields of statistical learning, even though it might be construed as more significant in smaller datasets. If one wants to avoid some specific biases, the solution is not to simply get a *bigger* corpus, one should have a *better curated* corpus (Bender and Friedman, 2018). To reduce the concept of bias to the quality of the data would however still be a mistake, the algorithms used themselves can give more or less importance to fringe cases and outliers, for example, and some tasks are implicitly conceived as being detrimental to some communities (Bender et al., 2021).

What does that say in our case? The position we choose to adopt here is that bias taken in a broader sense is not an issue for WE, it's a feature of all VSM. While that feature has proved to be a big issue for the bigger, performance-oriented models, especially when they are deployed

6 Should we choose to say writings are a kind of utterance.

at a larger scale, we argue that it can be useful from a data exploration perspective. Once we use WE to explore corpora, we want to keep the biases in word associations intact. From a general discourse/text analysis and humanities standpoint, the bias *is*, to an extent, the data we are looking for.

7.3.2 *Word embeddings for discourse analysis and stability*

Besides bias, a big issue that we might have is that, due to the stochastic nature of the algorithms used to construct WE, running a given algorithm several times on the same corpus can lead to very different results (an issue henceforth referred to as the *stability* of the embeddings). Both bias and stability can be serious issues for general-purpose language models, whose stated goals are higher performance in various tasks (such as those presented at SemEval). For the analysis of semantic variation across times or groups, we have seen that *bias* is not an issue to be solved, but a feature we want to analyze.

There is a technical issue that is much more problematic in our case and in applications of WE to discourse analysis in general. These methods are typically used with extremely large corpora containing billions of tokens and therefore cannot be used as such in many endeavors in the humanities and social sciences, where the size of the corpora used is not only limited, but sometimes also impossible to augment in a sensible way[7]. *Stability* quickly becomes an issue in these cases, as each iteration of the model can give very different results. If the quantitative measures we wanted to use to complete qualitative analyses happen to have so much variability that they are basically random, then they are no use at all.

Luckily, there have been attempts at solutions (see Antoniak and Mimno, 2018, applied in Rodman, 2020). We argue here that once the *stability* issue has been acknowledged and is taken into account to mitigate the observed results in word embeddings, they can be seen as a useful tool for discourse analysis.

These partial solutions include:

1. Fine-tuning
2. Bootstrapping

FINE-TUNING Fine-tuning (Howard and Ruder, 2018) a model is a fairly standard practice. The idea is to take a VSM trained on a very large corpus, possibly much larger than what we are actually working with, and use the matrix generated by that VSM as a starting point for a new iteration of the model over a smaller corpus. So instead of starting with random weights in the weight matrix $\mathbf{W_1}$, word2vec is initiated with weights that have already been trained, using a VSM that already has a meaningful structure. The idea here is that the larger corpus

7 For example, a study focusing on the works of one particular author cannot have a corpus that would be larger than the complete works of that author, which might turn out to be much smaller than the corpora generally used to create word embeddings.

is big enough for its vocabulary to be a superset of the vocabulary of the smaller corpus. The model trained on the larger corpus (or *pre-trained* model) is reliable in its standard predictions on word similarities, retraining it on the smaller corpus has the effect of biasing the VSM and end up on a model that is more representative of the distribution of words in the smaller corpus. This allows us to profit from the size and stability of the pre-trained model while having results in terms of data exploration that pertain to our own, smaller corpus.

Issues remain to be dealt with even with fine-tuning however.

BOOTSTRAPPING Bootstrapping is a classical technique for data augmentation that is used in many statistical endeavors. In proper statistics terms, *bootstrapping* is a *random sample with replacement*. The data augmentation via bootstrapping process consists in taking the set of all available data points (including repetitions) and choose one with uniform probability. The process is repeated until we reach the desired dataset size.

Why would we be interested in this? As stated in Antoniak and Mimno (2018), the presence of specific documents in a corpus can have significant effects on the cosine similarities between embedding vectors. Luckily, the paper also shows that one can produce reliable outputs from smaller corpora by controlling for the presence of specific documents and their lengths through bootstrapping of the corpora. In Antoniak and Mimno (2018), the corpus under study is bootstrapped at the document scale and new models are generated for each bootstrapped version of the corpus; the diverging model outputs can then be averaged, leading to stable results even for smaller corpora.

Following Rodman (2020), we have decided to use both fine-tuning and bootstrapping in the experiment we present in Chapter 8. The details of how we did this will be tackled there.

7.3.3 *Computing partial synonymy using word embeddings*

The idea then is the following: we suspect that some words in a given discourse have dogwhistley undertones, that they might mean something different in different audiences. What we should do is therefore gather a corpus of discourses on a similar issue made by people that have a similar ideology or are part of the same groups. We also need to gather a corpus of discourses on that issue that are produced by people from other groups. We then build a VSM for each of those corpus, using the same pre-trained model for both. To address the issue of stability, we train such a model for several bootstrapped versions of each corpus, and instead of directly using the values from our many VSM, we average the values within social groups.

Those averaged results are then compared across groups, the dogwhistley words should display a significantly stronger closeness (= higher cosine similarity) to words that we estimate could be synonymous to their implicit meaning in one group, and they should not display such proximity in the other, while non-dogwhistley words should not display such a difference. And that's it, right?

Not exactly. A final issue that remains is what Rodman (2020) calls "spatial noncomparability". The issue here is that the fine-tuning step in our work can lead to radical changes in the shape of our vector space. In short: we are not immune to the fact that the training algorithm might change the space in different ways according to its initial weights and to the corpus it is working on when trying to minimize the error function that serves as a medium for the building of the VSM. In our case, the pre-training ensures that we start with the same initial weights, however, given that our corpora and their bootstrapped versions can be rather small, it is possible that the presence of some documents drastically alter the general layout of the vector space (especially the longer documents).

This has one key consequence: the cosine similarity scores that we derive for the models are not necessarily directly comparable across groups. This has to be kept in mind when analyzing the outputs of the models. Thanks to the fine-tuning and averaging processes that our outputs undergo, the risk that the outputs are completely incomparable is mitigated, but it is still present. This is why, even though we have to compute the average cosine similarity scores, we do not use them in analyses to rather focus on the *ordering* of word similarities (which should not be impacted by spatial noncomparability in the same way).

We have attempted to use several techniques to compare the two VSM resulting from our computational enterprise, those techniques are detailed in Chapter 8.

7.4 CONCLUDING REMARKS

We have seen that a big limitation of the formal approach we have adopted so far is that the description that it resulted in fails to give a definition of dogwhistles that can be used as a working definition to distinguish what is from what is not a dogwhistle. For political, philosophical, and academic reasons, we might want to have such a definition. In this chapter we have made an argument for the use of computational methods, specifically methods based on WE to attempt such a definition:

> *Given a finite set of VSM* $\mathbb{E} = \{E_1, \ldots, E_k\}$*, with a specific group associated with each, a word is a dogwhistle in the distributional sense if it is the case that its nearest neighbors in* $E_i \in \mathbb{E}$ *are significantly different from its nearest neighbors in all* $E_j \in \mathbb{E}$ *with* E_i *different from* E_j*.*

Computational and quantitative methods are very obviously not the only methods that exist, and in this particular case they might not even be the most successful or intuitively appropriate, nonetheless, they provide ways to see what quantity and kind of information we can actually manage to extract from observable data while reducing the individual input of the scientist or experimenter. These methods applied to the phenomenon of dogwhistles might not necessarily give results that would be as satisfactory as careful qualitative analysis of the texts, but we still have faith that they can bring a new look on the matter, especially when the amount of data to be analyzed is just a little bit too big for a careful manual analysis.

We have tackled some of the issues that using methods like WE might bring, but all of this remains fairly abstract so far; those methodological considerations are applied directly to a study case in the next chapter, where hopefully all of it will seem a lot clearer.

In any case, what we are attempting to do here is, again, not to give a foolproof solution to the problem of dogwhistle detection. Rather, we are trying to see whether some computational tools whose use in this task seems philosophically motivated can actually be useful in this endeavor. In many ways, this is more of a test of the limits of available computational tools than it is a true detection task.

8

WE REPRESENTATION FOR DOGWHISTLE DISCOVERY

"VIZZINI:

He didn't fall? Inconceivable!

INIGO MONTOYA:

You keep using that word. I do not think it means what you think it means."

— Inigo commenting on how different people might understand words differently (*The Princess Bride*) (Reiner, 1987)

"[W]here I see one discourse, you may see a different discourse or no discourse at all. Our identification of particular discourses is going to be based on the discourses that we already (often unconsciously) live with"

— Baker (2008) quoted in Findlay (2017). This illustrates the previous point from *The Princess Bride* and applies it to the discourse analyst themself. (Findlay, 2017)

Chapter 7 presented arguments in favor of the use of WE for the exploration of corpora and a possible subsequent detection of dogwhistles. This chapter will show an attempt at doing just that. The detection of dogwhistles in corpora as we have discussed it requires us to have a corpus that includes not only utterances, but a fairly precise idea of what speakers believe. Obviously, we cannot have any access to speakers' internal monologues and hidden worldviews, but if we remember from Chapters 4 and 5, this is not about the actual intentions and worldviews of speakers but about what can be retrieved from their use of language along with their persona. Having information about the identity of the speaker and their general ideology should in theory be sufficient for listeners to make the necessary inferences and interpret dogwhistles.

So what we need is a corpus where in addition to utterances, we have access to information about the speaker that could allow us to place them somewhere in the political and personae space. We think we have found such a corpus with French Parliamentary debates. In the context of this work, we chose to focus on issues of gay rights, specifically on the issues of PACS (civil union) and same-sex marriage.

8.1 THE CORPORA

We will use here two distinct (though related) corpora. The first that we have used consists in the public debates at the French *Assemblée Nationale* surrounding the question of the legaliza-

tion of same-sex marriage. The second is similar in that it also consists in public debates at the Assemblée, but it is older and focuses on civil unions. We have chosen to add it *a posteriori* to see whether some diachronic analyses could be conducted. From now on, these two corpora will be respectively referred to as the MPT corpus[1] and the PACS corpus[2].

Why those corpora? A few reasons make them potentialy good candidates. Some of those reasons are related to material considerations:

- The content of the debates is freely accessible to anyone interested at `https://www.assemblee-nationale.fr`
- The subject matter at the scale of a given corpus is relatively homogeneous.
- The identity of the speakers in the debates is known and their positioning in the political space is known.
- The opinions of these people is known, at least insofar as both their discourse and their votes are available.

Beyond the material reasons that make those corpora good potential candidates, there are reasons to think that dogwhistles might have been present in these corpora; this is due both to the subject-matter being discussed and specificities of French politics.

8.1.1 *Gay rights at the* Assemblée Nationale

On the subject of secret signaling of the speakers' belonging to an ingroup while adressing a larger audience, the issue of same-sex marriage itself might be a promising example. It was very contentious at the time of the debates in France; in particular, religious conservative groups were very outspoken against the law and led many of the popular uprisings against it that occurred (Béraud et al., 2015; Coorebyter, 2013; Garbagnoli and Prearo, 2017; Louise, 2017; Théry and Portier, 2015). Due to historical, sociological, and political reasons (in particular due to the importance of the principle of *laïcité*, 'secularism'), French conservatives might avoid using religious terminology and argumentation, especially at the Assemblée Nationale. Conservative politicians were therefore facing the conundrum of wanting to appeal to their religious voters without being in the full capacity of using religious speech. We therefore hypothesize that they might use some secular words (like *nature* or *civilisation*) which, in their mouths and in the context of this debate, would acquire religious connotations.

Likewise, one recurring complaint against conservative politicians at the time was the use of the slippery slope argument according to which legalising same-sex marriage would lead

1 "*Mariage Pour Tous*", "Marriage For All", the name given in political discourse and the media to the set of legal texts that would open the institution of marriage to same-sex couples. The acronym actually emerged when used by opponents to these texts as (L)MPT, this time meaning "La Manif Pour Tous", "Protest for All".

2 "*PActe Civil de Solidarité*", "Civic Pact for Solidarity", the name chosen for civil unions in France, generally only referred to using the acronym.

to *PMA* ('In Vitro Fertilization') being available to more couples, which would in turn lead to *GPA* ('surrogacy'), which is illegal in France, to also become legal, see for example the following, taken from the MPT corpus:

(30)

a. **Philippe Meunier** (Against *mariage pour tous*) : Aujourd'hui le mariage, demain la PMA, et nous savons qu'au sein de la majorité certains souhaitent la GPA.
February 2nd, 2013
Marie-Georges Buffet (Pro *mariage pour tous*) : Par ailleurs, chers collègues de l'opposition, vous ne cessez de vouloir lier la PMA et la GPA, au nom de l'égalité.
(...)
Annie Genevard (Against *mariage pour tous*) : Nous pensons que la PMA constitue, avec la GPA, le véritable objectif des partisans du projet de loi.
February 3rd 2013

b. ***P.M.**: Today marriage, tomorrow IVF, and we know that some among the majority wish for surrogacy.*

***M.-G.B.**: By the way, dear colleagues from the opposition, you cannot refrain from linking IVF to surrogacy in the name of equality.*

(...)

***A.G.**: We think that IVF, along with surrogacy, constitutes the actual goal of the supporters of this text.*

We hypothesize that terms like *PMA*, *GPA* and *mariage* should be seen as related to each other. Because of other concerns at the time that politicians subtly presented the legalisation of same-sex marriage as equivalent to the legalisation of pedophilia, polygamy, zoophilia and other crimes and infractions[3], we hypothesize that terms such as *PMA* and *GPA*, and possibly even *mariage*, could also be associated with these other, more taboo, illegal practices.

Overall, the resulting situation is one in which it is possible that some words get imbued with supplementary meaning when used by, e. g., opponents to same-sex marriage. Various groups of listeners might then have various interpretations of a single message based on their knowledge about the political orientation of the speaker. It is therefore possible that a given word is used with different intended meanings according to the speaker's group. We hypothesize that such different meanings lead to different uses of the word and different collocations, following the basic ideas behind distributional semantics models and are therefore likely to lead to variation between our corpora.

3 According to SOS-Homophobie (2013), text available here: https://www.sos-homophobie.org/mariage-pour-tous-et-toutes/charte

corpus	**all**	**pro**	**anti**
tokens	624 483	169 525	375 630

Table 48: This table shows the number of tokens per corpus for the MPT corpus. These are the numbers after the corpus has been cleaned from numerical characters. The reason why the sum of tokens for the *pro* and the *anti* corpora does not add up to the number of tokens for the *all* corpus is that the many utterances produced by the presiding representative are omitted in the two position-specific corpora. The president's role in these debates is mostly to announce votes and give the floor to the next speaker, but they do not take part in the debates while they are on president duty, meaning their highly normalized discourse cannot be clearly defined as being *pro* or *anti* (although they do get to vote in the end).

8.1.1.1 *Same-sex marriage*

The MPT corpus consists of the debates surrounding the question of the legalisation of same-sex marriage at the French Assemblée Nationale. These discussions took place between January 29 and April 23 of the year 2013[4] and their transcript is freely available on the Assemblée Nationale's website[5]. There were 31 sessions discussing the issue at the Assemblée, the number of tokens for each corpus of interest is reported in Table 48. The entire corpus was annotated to show the identity of the speaker for each utterance as well as their political group and whether they supported same-sex marriage or not (based on both their discourse and the final votes); this annotation is detailed in Section 8.1.2. The corpus is very asymmetrical in that the representatives that were against same-sex marriage spoke a lot more (despite being a minority), leading to the three corpora shown in Table 48: the "*all*" corpus, containing all utterances by all Representatives; the "*pro*" corpus, containing only the utterances of Representatives in favor of the adoption of the same-sex marriage; the "*anti*" corpus, containing only the utterances of Representatives against the adoption of same-sex marriage.

8.1.1.2 *PACS*

The PACS corpus consists of the debates surrounding the institution of a civil union at the Assemblée Nationale. These discussions took place between November 3 1998 and October 13 1999[6]. Their transcript is fully available on the Assemblée's website[7]. There were 28 sessions discussing the issue, the number of tokens for each corpus of interest is reported in Table 49. The corpus has been annotated following the same standard as the MPT corpus; it is again fairly asymmetrical.

4 This includes a pause in the debates between February 12 and April 17, when a new version of the text was being written.

5 http://www.assemblee-nationale.fr/14/dossiers/mariage_personnes_meme_sexe.asp

6 This time including several pauses for the drafting of more versions of the text.

7 https://www.assemblee-nationale.fr/11/dossiers/pacs.asp

corpus	**all**	**pro**	**anti**
tokens	575 754	128 327	312 156

Table 49: This table shows the number of tokens per corpus for the PACS corpus. These are the numbers after the corpus has been cleaned from numerical characters. The same remarks hold regarding the total number of tokens.

8.1.2 *Annotation*

For the annotation of the corpus, we reproduced what was presented in Truan (2016), which is a flavor of TEI-XML annotation aimed at the annotation of public discourses. We have used the exact same standard as Truan (2016) with the exception that the `<floruit>` field in each participant info was changed from showing whether a speaker belongs to the opposition or not to showing the position regarding the issue at hand (same-sex marriage or civil union), with either "pro" or "anti"[8], based on the discourses of the representatives as well as on their vote regarding the texts.

In a few words, the corpus was separated into *utterances*, which are here understood as any string of words uttered by a speaker without interruption from other speakers. The speeches at the Assemblée Nationale are not completely spontaneous, most discourses are prepared in advance, but the responses by e. g., members of the government are not, and a lot of heckling takes place, sometimes interrupting the stream of discourse from a given speaker. For each utterance, an ID linking to the identity of the speaker is given, in cases the speaker is unidentified (for cases of heckling for example), the clerks in charge of transcribing the sessions usually mention the political affiliations of the speaker in question, determined by their position in the room. In those cases, the speaker was identified as their group's name. Here is an excerpt from the first session of the MPT corpus:

```
<u who="#BINET"> La très grande majorité d'entre elles
montrent que les enfants se portent ni mieux, ni moins
bien que dans les familles hétérosexuelles. Le nombre
de ces études devient considérable et impose un
faisceau de conclusions concordantes : les enfants
issus de familles homoparentales sont des enfants
comme les autres.
<incident>
```

8 A note on this: we are well aware that this erases a lot of the nuances that the debates brought, notably the many flavors of "anti" that one could find. One key distinction, for example, is that between the representatives opposing the law *per se*, with arguments revolving around the problems that, according to them, the law might bring, and those opposing the law on the grounds that there are more pressing things to discuss, like unemployment. That being said, a few things can comfort us in this choice. The first is that the distinction between the two groups is not as clear as one would think, with opponents just generally gleaning arguments against the policies; the second is that the representatives themselves belong to political parties and groups, and that on reforms such as those discussed here, the position of each representative is largely subsumed to that of the group, and these differentiations largely occur along group lines. Given that our annotation system does give us the group of each candidate, we could in theory retrieve some of that nuance.

```
    <desc>Applaudissements sur de nombreux bancs du
    groupe SRC.</desc>
</incident>
</u>

<u who="#MARITON"> Ce ne sont pas des études, mais des
manifestes !
</u>

<u who="#BINET"> Pour autant, les familles homoparentales
ne sont pas aujourd'hui des familles comme les autres, car
leur existence n'est pas reconnue par notre droit, leur
engagement de couple et la protection de leurs enfants sont
ignorés par la République. Dans ces familles, le divorce
n'existe pas et le juge ne peut départager un droit de garde
ni statuer sur une pension alimentaire ; dans ces familles,
le décès d'un seul des deux membres du couple peut créer des
orphelins,...
</u>

<u who="#MARITON">  Faux !
</u>
```

At the beginning of each transcription is a list of participants, which works in principle like a database about each. For all participants, the information about them is organized as such (taking the example of representative Hervé Mariton):

```
<person xml:id="MARITON">
    <persName>Hervé Mariton</persName>
    <sex>male</sex>
    <occupation>MP</occupation>
    <affiliation>Union pour un Mouvement Populaire</affiliation>
    <trait type="party"><desc>Conservative </desc>
    </trait>
    <floruit>anti</floruit>
    <residence>Drôme (3ème circonscription)</residence>
    <nationality>French</nationality>
</person>
```

Each entry thus contains the ID of the speaker, their full name, their sex[9], their occupation (*MP, Président.e de l'Assemblée, Rapporteur.e, etc.*), their party affiliation, the general political family of their party, their position on the subject at hand, their constituency and their nationality[10]. Of all this info, we mostly use the `<floruit>` information, although in principle we could hypothesize on the importance of others, notably the party affiliation.

8.2 TRAINING EMBEDDINGS

If we look at computational research in discourse analysis, the main tool used by researchers has been *keyword analysis* (Scott et al., 2001). In a few words, the idea behind keyword analysis is to give each word in a corpus a *keyness factor* to identify those words that could be seen as keywords. *Keywords* are words used significantly more often in some texts than in others, and they are discovered through comparing relative frequencies of words across corpora, either using a reference corpus (like, e. g., the BNC) as a means of comparison, or comparing both corpora directly to each other. For each word in each corpus, their actual frequency is compared with an expected frequency computed using the two corpora's data. These values are then used to compute the *keyness factor* for each word in each corpus, typically using a χ^2 or log-likelihood metric. Ranking words by decreasing order of keyness in a corpus therefore allows one to see which words appear comparatively more in one corpus rather than in another, acting like "lexical signposts, revealing what producers of a text have chosen to focus on" (Baker, 2004, p. 90).

Regarding LGBT+ rights in the context of the UK for example, such measures have been used in discourse analysis works focusing on the identification of ideological differences in arguments for/against lowering the age of consent for gay sex (Baker, 2004); civil partnerships (Bachmann, 2011) or same-sex marriage (Findlay, 2017). In these cases, the corpora under study are divided between *supporters'* and *opponents'* utterances, and these corpora are in turn compared using keyword analysis (with or without resorting to a reference corpus).

The keyword method is very useful in uncovering discourses in contexts where different groups will tackle different topics and therefore use a different lexicon. However, in cases where the groups in question use the same lexicon with a similar frequency, they cannot bring much information. Similarly, they cannot be used to compare how a given word is used by different groups when there is no major difference in the number of times that word is uttered by either group. In other words, the keywords method can give us information on which words are favored in a given group compared with another, but they cannot tell us how these words are used, what they *mean* for the group in question. In the case of dogwhistle study, this is a limitation that one might want to address.

9 This information is not useful in that case, we kept it to remain consistent with what is used in Truan (2016).
10 Our corpus contains exclusively French people, the original annotation schema presented in Truan (2016) was used to compare discourse variation between various members of the EU Parliament.

But before that let's consider for example the French context, where some discourses lack the "lexical signposts" that keywords allow us to uncover. In the case of the gay marriage debates that took place at the Assemblée Nationale between January 29th, 2013 and April 23rd, 2013, both sides of the debate (henceforth *pro* and *anti*) argued their views were in line with the Republican values of *liberté, égalité* and *fraternité*, as shown in (31) and (32):

(31) a. **Christiane Taubira** (Pro *mariage pour tous*) : Nous disons que le mariage ouvert aux couples de même sexe illustre bien la devise de la République. Il illustre la liberté de se choisir, la liberté de décider de vivre ensemble.
Yves Fromion (Against *mariage pour tous*) : Et la liberté des enfants d'avoir un père et une mère ?
Christiane Taubira: Nous proclamons par ce texte l'égalité de tous les couples, de toutes les familles.
Pierre Lequiller (Against mariage pour tous) : Et les enfants ?
Christiane Taubira : Enfin, nous disons aussi qu'il y a dans cet acte une démarche de fraternité, parce qu'aucune différence ne peut servir de prétexte à des discriminations d'État.
January 29th, 2013

b. *C.T.: We say that opening marriage to same-sex couples illustrates the Republic's motto. It illustrates the liberty of choice, the liberty of deciding to live together.*
Y.F.: What about the liberty for kids to have both a father and a mother?
C.T.: With this text, we proclaim the equality of all couples and of all families.
P.L.: What about the kids?
C.T: Finally, we also say that this here is another step towards fraternity, because no difference can serve as an excuse for State discrimination.

(32) a. **Hervé Mariton** (Against mariage pour tous): Pour que la liberté soit aussi responsabilité, pour que l'égalité soit aussi respect de la différence, et pour que la fraternité se fonde, plutôt que sur la division, sur l'unité, nous ne voterons pas ce texte.
January 29th, 2013

b. *H.M.: For liberty to also be responsibility, for equality to also be the respect of difference, and for fraternity to be founded upon unity rather than division, we will not vote in favor of this text.*

While keywords approaches can help us determine the various themes that are tackled by one group or another on a given question, word embeddings allow us to specifically see how some words are used, and their semantic associations. In the sense that they allow us to ask very different questions, the two methods are complementary.

We will here focus on the following questions:

- Can word embeddings bring useful information for synchronic semantic analysis across groups?
- How can we address the issue of corpus size when using word embeddings?
- How can this tool be used to uncover discourses in French gay marriage debates and how does it compare to more traditional computational approaches (such as keywords)?

In the context of this work, we have used the word2vec approach, first presented in Mikolov et al. (2013). Word2vec is a shallow, two-layer neural network which takes as input a large corpus and outputs a vector-space of several hundred dimensions in which each unique word of the corpus is assigned a vector. In this work, we specifically used the CBOW (Continuous Bag-of-Words) algorithm, which works as described above, by trying to predict a target word from its surrounding context[11].

Word2vec and similar algorithms have notably been used to track the meaning associations of words over time with interesting results (Garg et al., 2018). In our case, we would like to compare word meaning associations not over time, but across ideologically-opposed communities. The task is similar: we have a corpus that we split into smaller corpora depending on political orientation. An issue that we face, which does not apply to Garg et al. (2018) and others, is corpus size. The corpora used in the works cited above are large corpora, and in fact the corpora used to train algorithms like word2vec are typically a lot larger than the corpus we have here. This can lead to several issues: first, it is possible that the word similarity results that we get are not very precise, especially for less frequent words; second, because word2vec is initiated using random weights, two iterations of the algorithm do not necessarily produce identical results. Because the corpus is smaller, it can be the case that these results differ immensely. Those issues are discussed a bit more in Section 7.3.1.

This is not ideal. Luckily, there are solutions to mitigate these issues and ensure some stability of our results (or at least measure how certain we can be about the output). These solutions are notably described in Antoniak and Mimno (2018), and then applied in Rodman (2020).

Following Rodman (2020), we have decided to both fine-tune the model to our corpora and to bootstrap them and generate several models, the outputs of which we then averaged to obtain the results presented in section 8.3.2. The pre-trained model is one of those found in Fauconnier (2015); it consists of 500-dimensions vectors and was trained on a lemmatized version of the frWac corpus. The frWac corpus (Baroni et al., 2009) is a 1.3 billion word web-crawled French corpus. Because of the origin of the documents it contains (various websites of the `.fr` domain), it is likely to contain language that is different from the language that can be found in our own corpora (it sounds very likely that speakers at the Assemblée Nationale

11 The other algorithm, *skip-gram*, predicts the surrounding context given a target word. Both algorithms could be used here, we preferred CBOW because it is computationally lighter than skip-gram. It tends, however, to smooth the context and although it gives accurate predictions for more frequent words, less frequent words' results are more erratic than they would be using skip-gram. More information about everything word2vec related can be found in Appendix C.

do not talk like users on internet forums). The corpus was also constructed a few years before these debates were at the center of attention in France. Nevertheless, because of its size, it gives a solid base for the meaning of most common words found in our own corpora.

For the bootstrapping phase, the question arose of choosing which units in the corpus we wanted to bootstrap. Rodman (2020) has a corpus of many smaller documents, the most appropriate unit for us was to use *utterances* as the equivalent to Rodman (2020) "documents". The reasoning behind this is that each intervention by a representative at the Assemblée is supposed to be self-contained (the speeches are prepared in advance), whereas the sessions themselves contain many utterances each and are a lot more variable. Because some of the interruptions to utterances sometimes force the speakers to address their audience before resuming with a linguistic blank slate, the two parts of a discontinuous intervention were treated as separate utterances, as presented in Section 8.1.2. Once we have generated a model from each of the bootstrapped versions of our data, we generate the lists of closest semantic neighbours by computing the mean cosine similarity between words across all the models generated, along with its confidence interval (following Antoniak and Mimno, 2018).

8.3 LOOKING AT THE CORPORA

Before seeing any results, let's have a reminder of what we hypothesize to witness here:

1. Differences in broad terms often used in political discourse (*liberté, égalité, fraternité*... not unlike what we saw in Section 6.3.2).
2. Differences in usage/(distributional) meaning for words pertaining to the lexicon of the issue at hand (*mariage, PACS, PMA*...)
3. Religious undertones on seemingly non-religious terms (Théry and Portier, 2015), the most reminding of dogwhistles.

We can first of all try to see what a keywords approach can give us in terms of information. The keywords analysis and computations was done using the AntConc software (*AntConc (Version 3.5.9) [Computer Software]*).

8.3.1 *Keyword analysis*

The keyword metric that we use here is the log-likelihood from Rayson and Garside (2000), defined as:

$$2 \sum_i O_i \ln \left(\frac{O_i}{E_i} \right)$$

Where O stands for the observed counts of a word and E its expected value given the target and reference corpora, computed as follows:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

Where N values are the total word count for each corpus[12]. Using this metric, any keyness score that we obtain can be considered to result from a significant difference (95^{th} percentile, or $p = 0.05$) when it is higher than 3.84.

Besides the choice of statistics[13], the other main parameter that we can influence is the reference corpus. Geluso and Hirch (2019) underlines that this choice is of utmost importance, as different reference corpora will give very different results. Specifically, more fine-grained analysis requires more similar reference corpora. In the present case, we are looking for qualitatively different arguments in very specific corpora, so we want to have a reference corpus that is as close as possible to the target corpora. Our target corpora are all our sub-corpora, meaning both the *pro* and *anti* sub-corpora in both the MPT and PACS corpora. This means that we have a total of 4 target corpora that we would want to analyze, and they are all very similar. One simple solution to the reference corpus problem is therefore simply to take the concatenation of all sub-corpora as a reference corpus: it is resolutely larger than any individual target corpus; there are no words that appear in the target corpus without ever appearing in the reference corpus[14]. In this case, the information we have is basically: "when compared to the entirety of the debates, the keywords that are most representative of the $\frac{\text{pro}}{\text{anti}}$ arguments regarding $\frac{\text{MPT}}{\text{PACS}}$ are the following" results are presented in Tables 50 and 51. The process was done using the lemmatized versions of the corpora (the lemmatization step was done using TreeTagger, Schmid, 1994), the reason behind it is that we are interested in topic analysis and the general declension phenomena are not thought to be interesting or carry meaning in that case.

In spite of the noise attributable to lemmatization/tokenization issues, the tables can already show us some interesting patterns. In the case of the MPT corpus, some of the different arguments put forward by each side of the debate clearly transpire from the keywords. For example, the *pro* corpus shows that words like "famille", "homoparentales", or "égalité" are apprently representative of the content of that side of the debate. This is in keeping with the idea that the proponents of the law focused a lot on the idea of giving a legal foothold to families with same-sex parents that already existed at the time, with the idea of giving them the same rights as heterosexual married families with regards to adoption and to authority over the children, especially in case of separation. In the *anti* corpus, by contrast, we see that words like "père", "mère" or "filiation" underline the argument stating that all children should have a right to both a mother and a father to ensure that their mental development

12 A step-by-step explanation of the log-likelihood calculation can be found here: http://ucrel.lancs.ac.uk/llwizard.html

13 The choice we have made here is standard, but there are other relevant statistics, see e. g., Pojanapunya and Todd (2018).

14 This might lead to unwanted cases where a word that is in fact fairly common and uninteresting, but just so happens to never appear in the reference corpus would be thought of as having a very high keyness score.

"pro"		"anti"	
KEYWORD	KEYNESS	KEYWORD	KEYNESS
l	955.53	l	1643.41
d	561.11	d	1204.92
n	492.02	qu	747.46
c	467.35	c	702.00
qu	375.56	n	655.31
défavorable	231.80	nous	351.97
s	167.06	mère	312.97
j	166.71	père	275.16
famille	165.46	s	253.80
enfant	164.42	j	248.60
mariage	144.38	vous	196.65
homoparentales	119.70	enfant	164.82
parent	104.83	projet	153.66
égalité	103.43	notre	143.41
couple	98.67	filiation	135.22
adoption	96.48	gpa	128.75
avoir	92.35	femme	125.84
même	86.97	balai	123.09
état	83.27	adoption	121.15
parents	80.25	état	111.74

Table 50: Top 20 keywords for the sub-corpora in the MPT corpus. The lemmatization and tokenization has brought a lot of noise/errors (notably regarding determiners and other function words, but not exclusively), we can still see that some interesting nouns emerge in both corpora.

"pro"		"anti"	
KEYWORD	KEYNESS	KEYWORD	KEYNESS
défavorable	1150.55	pacs	1244.12
ne	390.75	ne	642.58
pacs	359.32	que	480.44
repousser	340.30	se	348.22
ce	186.62	de	287.92
le	152.92	avantage	230.27
que	143.38	le	217.50
la	140.65	contrat	206.09
se	138.64	la	182.26
vie	122.04	fiscal	175.02
commission	117.99	pacsés	145.21
solidarité	106.32	ce	133.56
de	104.28	problème	118.72
lecture	103.99	plus	117.28
concubinage	100.03	rupture	115.84
marier	79.82	vie	104.59
avoir	78.66	un	102.45
proposition	73.52	en	99.36
droit	70.24	statut	96.19
déjà	66.33	on	91.82

Table 51: Top 20 keywords for the sub-corpora in the PACS corpus.

went well. We also see the acronym "GPA", in keeping with the slippery slope argument that we have discussed earlierd.

If we look at the results for the PACS sub-corpora, there are also interesting points to be made, especially on the *anti* side of things, where the word "fiscal" underlines the general suspicion that PACS itself would be used to take advantage of the broadening of the notion of *tax household* and be generally done in bad faith in the hope of circumventing the system. The *pro* side on the other hand uses notably the word *concubinage*, which is the French legal term for the case when two people constitute a *de facto* household, possibly with children, but are not married. At the time already, the notion of families with same-sex parents was discussed. The goal of the proponents of PACS was, among other things, to give a legal foothold to families with same-sex parents.

8.3.2 *Word embeddings*

As stated earlier, keyword analysis does not allow us to have information on the alleged meaning difference in words across corpora, which WE supposedly can give us. Before generating the embeddings, the corpus was automatically lemmatized, in part to correspond to the pre-trained embeddings we had chosen, but also because general declension phenomena, like plural markings, were not particularly interesting to us here.

As stated earlier, we proceeded to generate bootstrapped versions of our corpora (100 of them) before tuning the pre-trained embeddings on those bootstrapped versions as well as the original. Each bootstrapped version led to a fine-tuned model, and all measures that we now present are the *average* measure over all of the models, for a given corpus/sub-corpus. All of this is illustrated in Figure 12.

As described in Figure 12, all the cosine similarity measures that we compute on our models are then averaged, and it is that average measure that we will use, although as stated in Section 7.3.1, we will not use the resulting average measure directly; we will only use ranked word similarities, and not the similarity measures themselves because of spatial non-comparability. The differences we expect to observe are thus to be found in the *lists of most similar words* to any target term. But how will we account for that difference? There are several things to be considered here, and the most prominent one is *control*.

We have first established a list of words for which we expect to see no clear difference between *pro* and *anti* discourse. The words we chose for this test were words which we assumed were used similarly by both sides of the debate. These words include deictic words (*ici, hier...* here, yesterday...) as well as words specific to the legal sociolect used at the Assemblée (*séance, amendement, article...* session, amendment, article...).

While some difference is to be expected due to the stochastic nature of the process, when we compare the closest semantic neighbours to these words in the *pro* and then the *anti* corpus, this difference should not be as great as with words we suspect are used differently. We assess

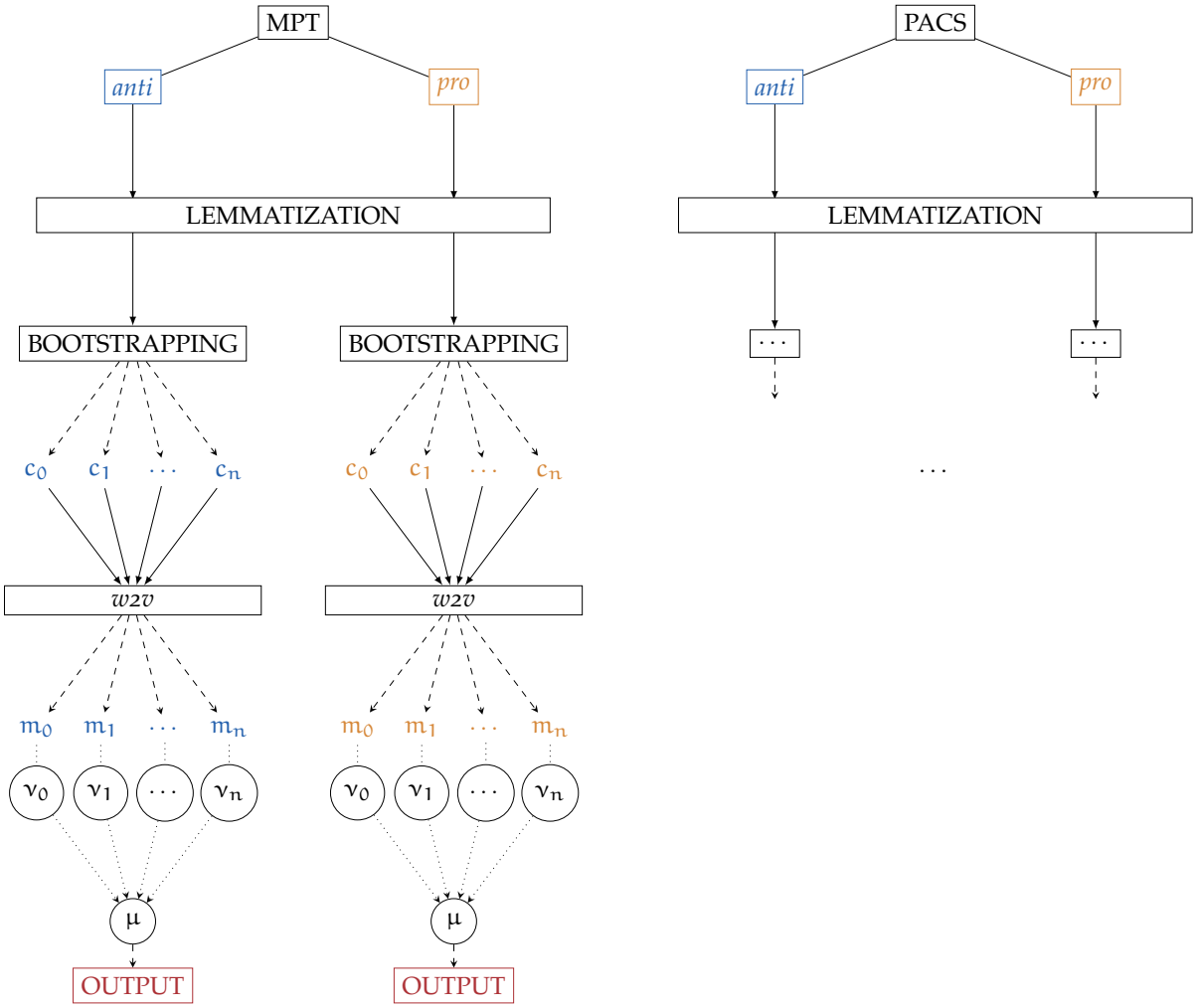


Figure 12: Overview of the processing pipeline leading to the measures that we use for analysis. Each bootstrapped corpus is used to generate a word2vec model, the measure ν that we are interested in (e. g., cosine similarity) is computed on each model, and we then compute the μ average of those measures. μ is the value that we then work with. The process is the same for the MPT and PACS corpora and their sub-corpora.

the differences by counting how many words among the 12[15] closest semantic neighbours are identical across models. This simple measure is therefore read as follows: 12 is the maximal score and means that all 12 closest semantic neighbors are different across corpora for the word of interest. 0 means the lists are identical.

We will see that this differentiation metric was not sophisticated enough to give us any valuable information and did not allow us to make a clear difference between test words and control words. This is due in part to the fact that this measure does not take *order* into account, just the content of the 12 most similar words list. We have therefore tried to find a measure that could take into account the order in which most similar words appear while not limiting itself to the first 12. The simple count of differences would have only worked with very big and obvious differences.

The second measure relies on the computation of *Spearman's* ρ. It works as follows: we take the intersection of the vocabularies of the two model families (*pro* and *anti*), we then use the ordering of similarity for all words in the *pro* models, with each word a label is associated, giving its position in the list; we then compare it with the ordering of words in the *anti* models. Using the same label-number pairing as the *pro* models, we end up with two ordered lists of numbers, one for each model family. We then use Spearman's ρ to check how different the ordering of the two lists are. These measurements do not appear to show a clear distinction between control words and test words either, it turns out that the fine-tuning of the model brings along a great number of changes also in the least similar words.

Though inconclusive, we have presented the results of computing these metrics in the following sections, where we take a direct look at the results for both the MPT and PACS corpora.

8.3.2.1 *MPT*

The control words along with their differentiation metrics across *pro* and *anti* corpora (both the count metric and the Spearman's ρ metric) can be found in Table 52. The words we investigated expecting to find interesting differences can be found in Table 53 (again with the differentiation metrics). These words were chosen following intuitions that we had from reading the corpus itself. As mentioned earlier, the values of "*liberté*", "*égalité*", and "*fraternité*" are all invoked to defend different positions. The words "*nature*" and "*naturel*" were also tested, as their use has been commented on in Théry and Portier (2015), seemingly indicating a rupture between secular discourses and Catholic conservative discourses. "*Mariage*" and "*famille*" were also used, given that the debates at their core questioned their definitions. We can also add the word "*adoption*", because adoption was also extended to same-sex couples with this text. Finally, "*GPA*" and "*PMA*" were studied because of the slippery slope argument put forward by the opposition.

All the words presented in Tables 52 and 53 were already present in the vocabulary of the pre-trained model.

15 12 was chosen arbitrarily, because it made the subsequent figures more pleasing to the eye.

word	**difference**	ρ
amendement	7	0.02754716
article	7	0.001340266
assemblée	3	0.01551367
député	3	0.0954906
hier	6	0.1340674
ici	7	0.06237997
loi	6	0.01623462
président	7	0.05851079
rapporteur	3	−0.01754226
vote	1	0.04566755

Table 52: These are the control words used in our study along with the difference measurements between *pro* and *anti* models for the MPT corpus. The central column is the initial count-based difference measurement between lists of closest semantic neighbours. 12 is the biggest possible score (meaning all 12 closest semantic neighbours are different between *pro* and *anti*). The rightmost column is the Spearman's ρ score that was computed when attempting to do the more in-depth comparison mentioned in section 8.3.2.

word	**difference**	ρ
mariage	7	0.04984883
nature	5	0.0416876
naturel	4	0.08985369
famille	8	0.05748241
GPA	8	0.05427596
PMA	11	0.06521176
liberté	4	0.1114692
égalité	3	0.02781825
fraternité	1	0.1005997
adoption	9	0.007430747

Table 53: These are the test words used in our study along with the difference measurements between *pro* and *anti* models for the MPT corpus. The column layout is the same as that in Table 52

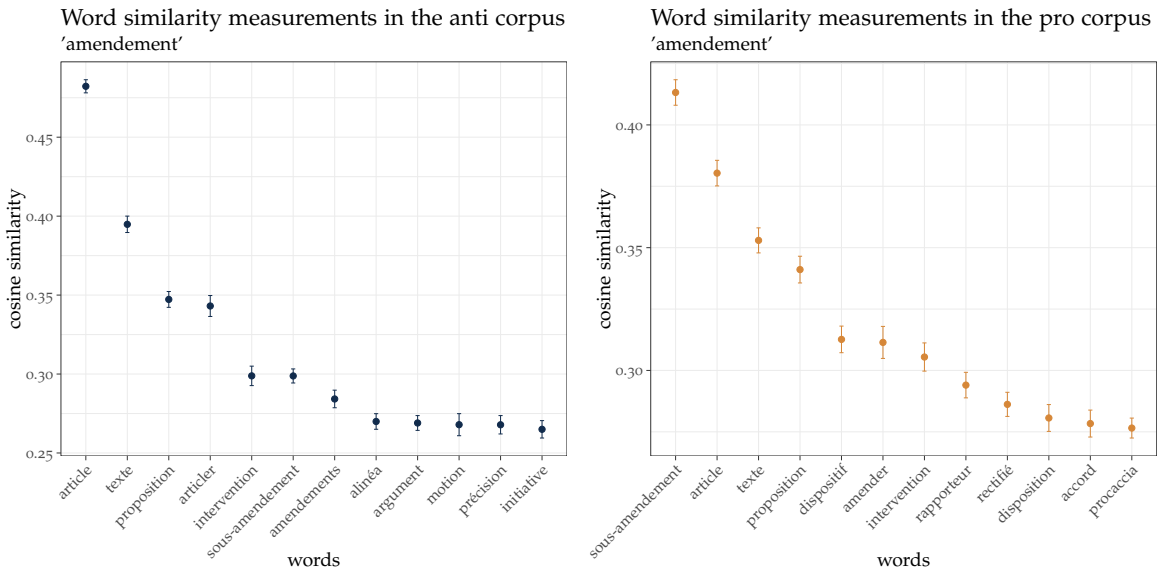


Figure 13: Comparison of closest semantic neighbours for *amendement* across MPT models. The error bars are 95% confidence intervals, showing that our methods have led to reasonably stable similarity orderings. y axes show the mean cosine similarity across models between *amendement* and the words on the x axes. The models trained on the *anti* corpus are on the left, the ones trained on the *pro* corpus are on the right.

We can see that the differentiation measure that we have put forward does not allow us to find a way to differentiate between test words and control words, it seems that the fine-tuning process brings with it a lot of changes in the VSM, for both most similar words and least similar words.

That being said, we can still look in more detail at what the differences actually are. Figures 13 and 14 can show us a clearer picture by comparing these lists of closest semantic neighbours directly. If we compare two words that obtained similar difference scores, like *mariage* for the test words and *amendement* for the control words, we can see that although they are rated similarly, the differences they display are not necessarily similar. In the case of *amendement*, the closest semantic neighbours across corpora are not the same, but they mostly belong to the same subset of specialized lexicon. This is also what is observed for the rest of the control words[16]. Regarding *mariage*, however, the differences are more interesting in terms of discourse. For example, the word *mariage* in the *anti* corpus has among its closest semantic neighbours the word *filiation*, in keeping with the conservative idea that marriage is conceived first and foremost through the lens of procreation.

The word *adoption* has *homoparentalité* ('homoparentality') as closest semantic neighbour in the *anti* corpus, and indeed adoption by same-sex couples is a key issue in the *anti* discourse. The word *famille* is more similar to *société* than it is to *familial* in the *anti* corpus, whereas the word *société* is not among the 12 closest semantic neighbours in the *pro* corpus, underlying again the conservative ideal of society being built on the traditional family.

16 With the notable exceptions of *rapporteur*, where we find a number of proper nouns, and *ici*, which is mostly associated with website interfaces.

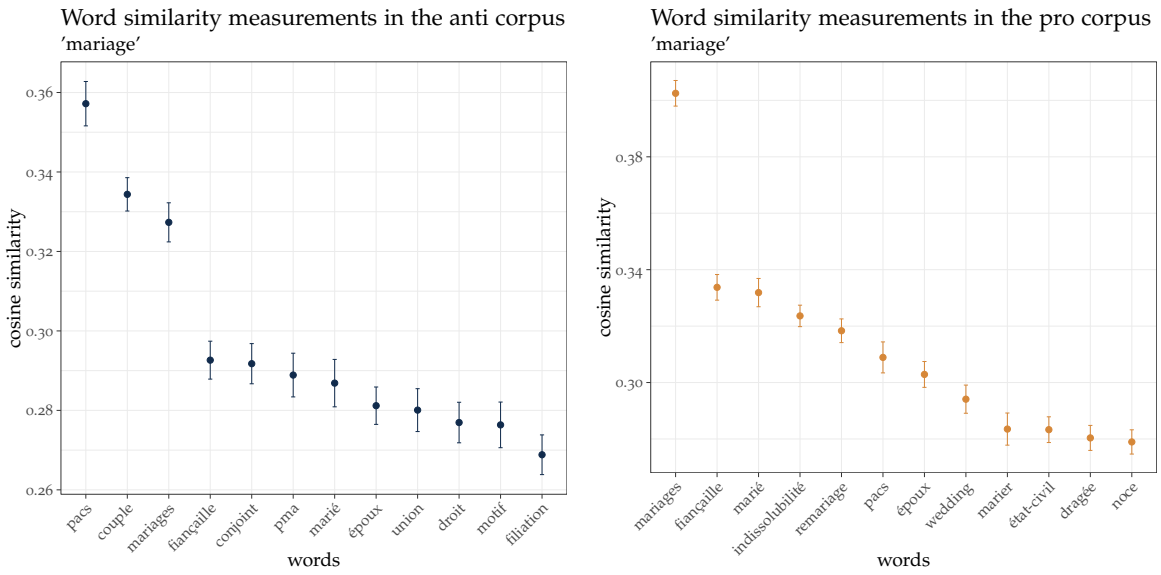


Figure 14: Comparison of closest semantic neighbours for *mariage* across MPT models. Overall layout identical to that of Figure 13.

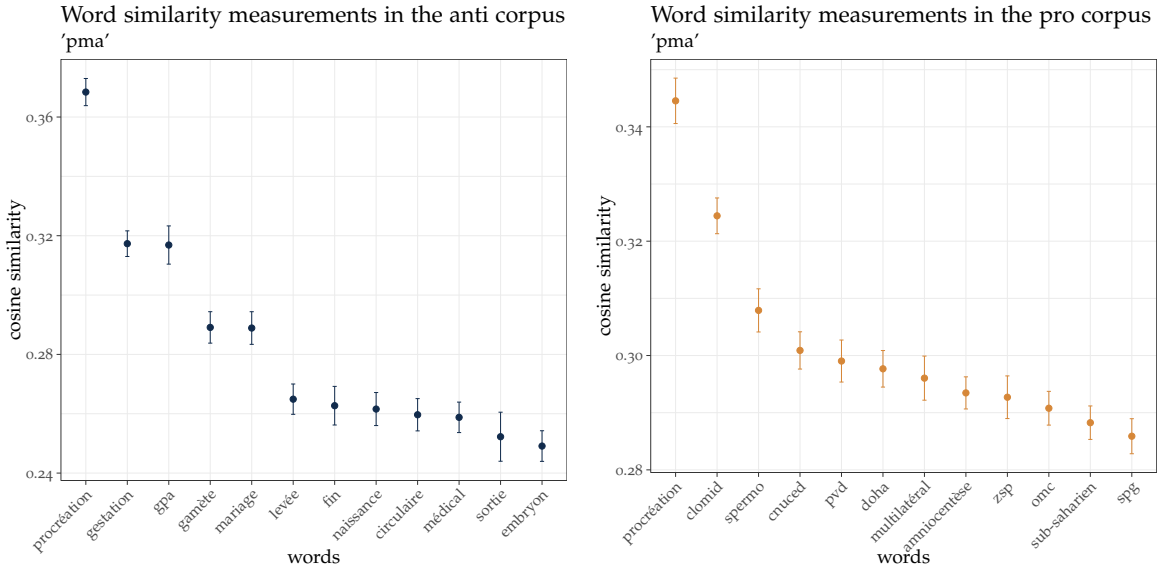


Figure 15: Comparison of closest semantic neighbours for *PMA* across MPT models. Overall layout identical to that of Figure 13.

Regarding *PMA* and *GPA*, we can observe the interesting fact that while *PMA* appears as one of the closest semantic neighbours to *GPA* in both corpora, that relation is not symmetrical. The word *GPA* is not among the closest semantic neighbours to *PMA* in the *pro* corpus, whereas it is the third closest semantic neighbour to *PMA* in the *anti* corpus, in keeping with the confusion that was maintained by conservatives between the two (Figures 15 and 16). See Appendix D for more comparisons.

What we would need is a measure that allows us to know not *how different* the word similarity rankings are, but *how interestingly different* they are. The difference between lists is a given, what we want is to have an estimation of the *degree* of difference, its "unexpectedness". An attempt is presented in Section 8.3.3.

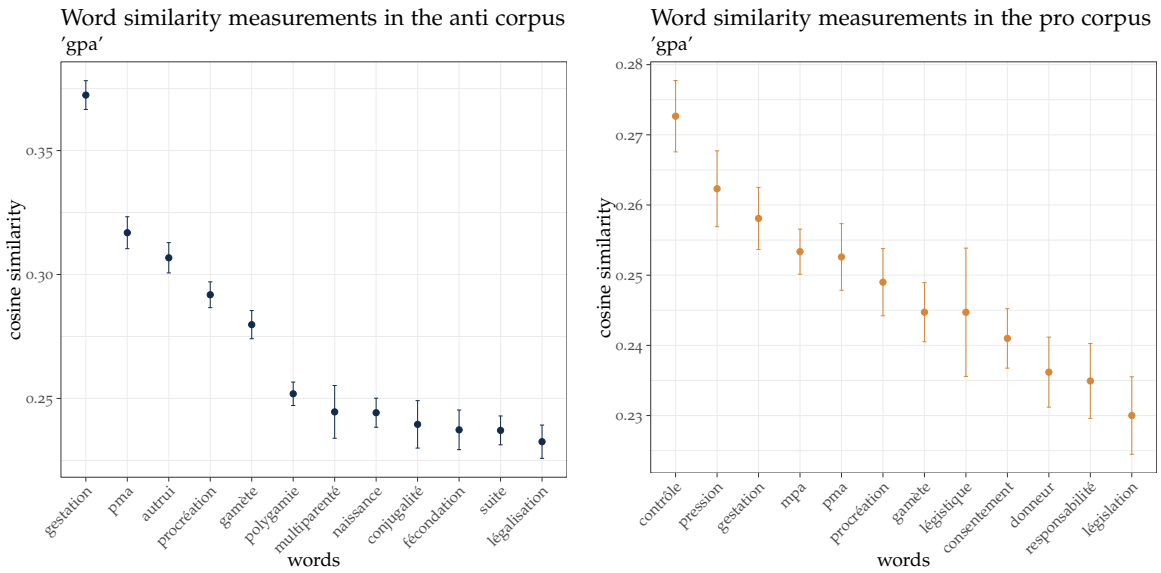


Figure 16: Comparison of closest semantic neighbours for *GPA* across MPT models. Overall layout identical to that of Figure 13.

While the purely quantitative approach does not allow us to isolate the interesting differences in distribution in our corpora, these techniques can still allow us to explore the data in new ways that can complement a qualitative review of discourses. Ultimately, the techniques applied to circumvent the issue of corpus size did not lead to the stability expected to conduct systematic quantitative analysis of the data, but the results obtained still lend themselves to interesting qualitative analyses going beyond the kind of conclusions that can be reached using more standard approaches to discourse analysis.

8.3.2.2 *PACS*

We will now reproduce the last section in the context of the PACS corpus. In this case, we were initially only interested by the corpus from a diachronic point of view, and wanted to see whether the test words we had looked at in the MPT corpus were used differently then. Upon closer analysis of the corpus, however, the same themes emerged then, and the arguments presented were very similar. It is likely that looking for test words specific to the PACS corpus, we would have chosen fairly similar ones to those that we chose with the MPT corpus in mind, if not the exact same words.

The original difference measurements per test and control word are presented in Tables 54 and 55

Similar to the MPT corpus, these metrics are not very convincing (if anything, the test words are much more similar across corpora this time around). The corresponding plots for "amendement", "mariage", "PMA", and "GPA" are found in Figures 17, 18, 19 and 20, respectively.

There are a few things to discuss here, from a diachronic point of view. First of all, we note one thing: "GPA" is not a part of the PACS corpus! What can we say about this?

word	**difference**	ρ
amendement	6	−0.03321359
article	8	−0.02134574
assemblée	7	0.04942737
député	3	0.1084174
hier	5	0.1540991
ici	5	0.0677787
loi	5	0.02025849
président	4	0.09541387
rapporteur	6	−0.02631045
vote	1	0.05588607

Table 54: These are the control words used in our study along with the difference measurements between *pro* and *anti* models for the PACS corpus. The column layout is the same as that in Table 52.

word	**difference**	ρ
mariage	4	0.05925414
nature	6	0.05472165
naturel	2	0.1145404
famille	5	0.1050941
GPA	0	−0.09938478
PMA	1	0.06935004
liberté	4	0.1065231
égalité	1	0.03664677
fraternité	1	0.1144403
adoption	5	0.02352368

Table 55: These are the test words used in our study along with the difference measurements between *pro* and *anti* models for the PACS corpus. The column layout is the same as that in Table 52. The reason why the counts for "GPA" are identical is tackled in Section 8.3.2.2.

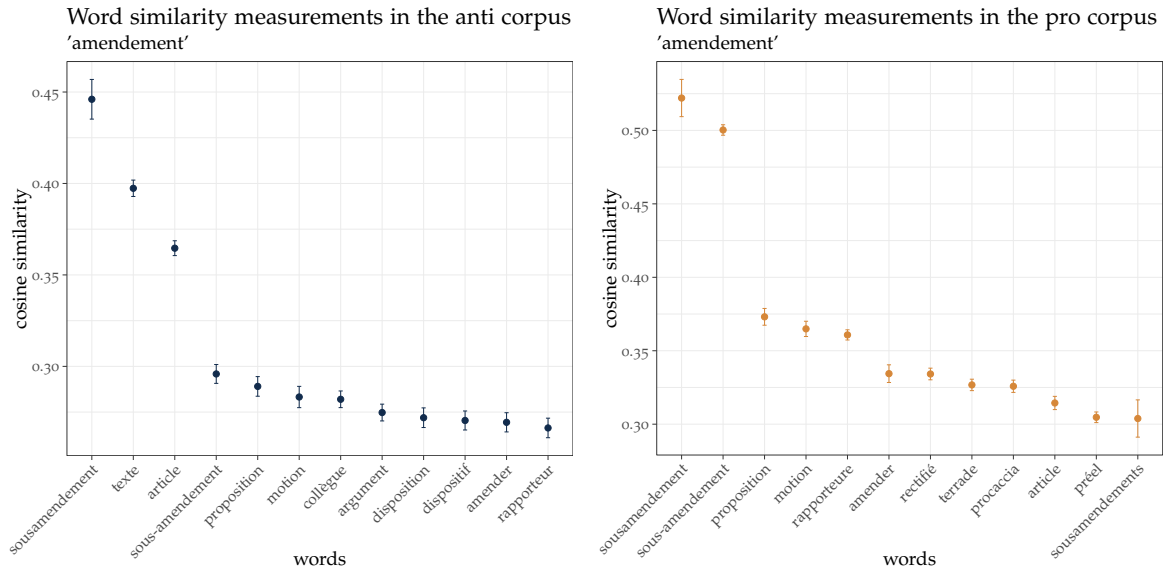


Figure 17: Comparison of closest semantic neighbours for *amendement* across PACS models. Overall layout identical to that of Figure 13.

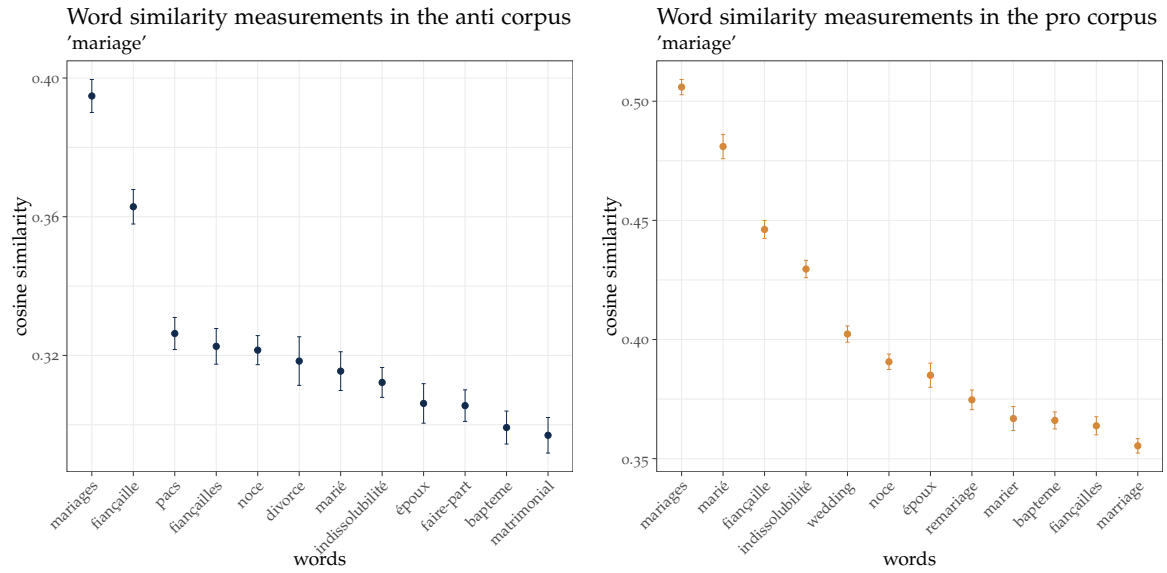


Figure 18: Comparison of closest semantic neighbours for *mariage* across PACS models. Overall layout identical to that of Figure 13.

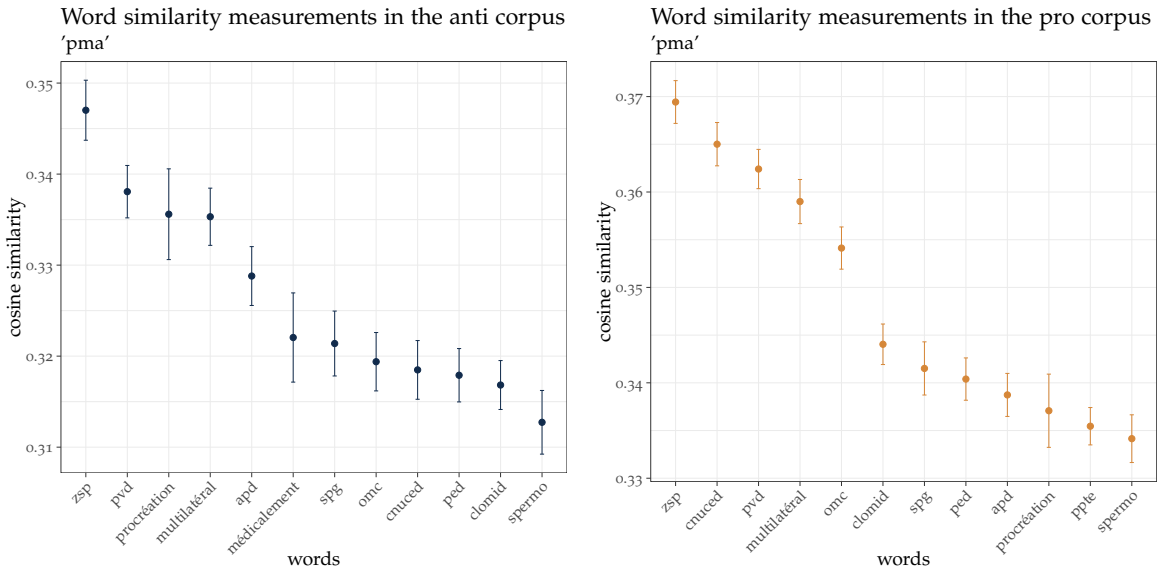


Figure 19: Comparison of closest semantic neighbours for *PMA* across PACS models. Overall layout identical to that of Figure 13.

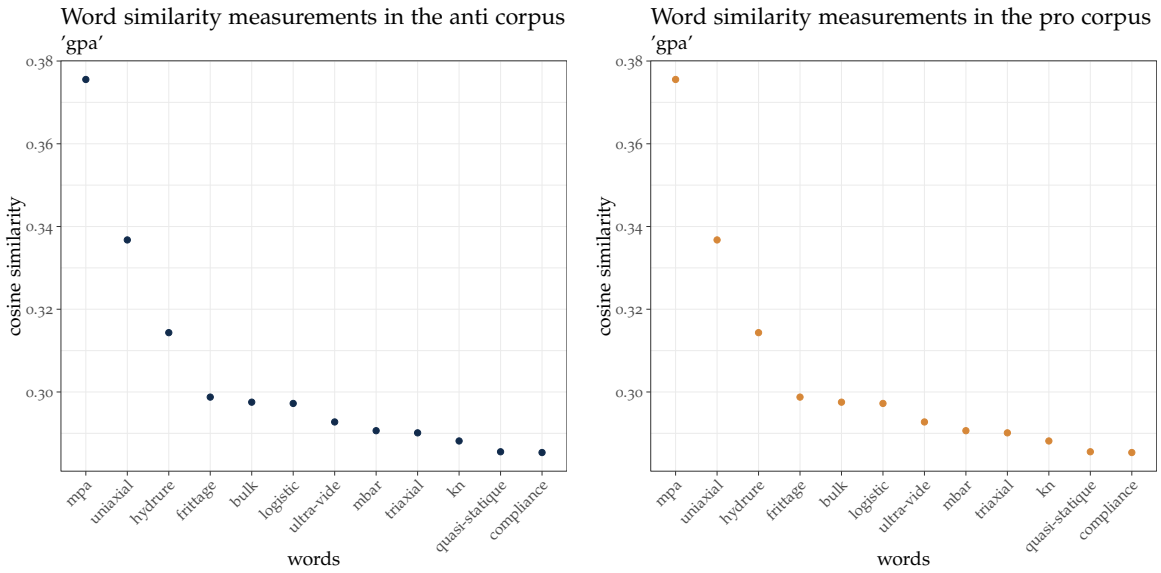


Figure 20: Comparison of closest semantic neighbours for *GPA* across PACS models. Overall layout identical to that of Figure 13.

- The concept of surrogacy is very ancient[17], so this does not mean that the concept of *gestation pour autrui* was not valid at that point, most likely it just means that the acronym was not used in this sense. Looking at the corpus, we find few mentions of the issue, under the expression "mère porteuse" (*surrogate mother*), but the debate at the time was much more about the broadening of the concept of adoption, which, although it was not illegal for homosexual individuals, was largely impeded, see (33).

(33) a. **Christine Boutin** (Against *PACS*) : [L]'adoption est ouverte aussi bien aux couples mariés qu'aux personnes seules âgées de plus de vingt-huit ans [...] l'adoption d'un enfant par un couple homosexuel, si elle n'est pas interdite par la loi, est limitée par un arrêt du Conseil d'Etat du 9octobre 1996, qui refuse de « donner le droit à l'adoption à un homosexuel » [...] le bénéfice de l'insémination artificielle ou d'une PMA est accordée exclusivement aux couples mariés ou aux concubins qui vivent ensemble depuis deux ans. Quant au recours à une mère porteuse, elle est évidemment interdite par la loi [...]. Officiellement, les promoteurs du PACS se refusent à lui donner toute possibilité d'influer sur l'état des personnes et renvoient ces enjeux à des lois futures sur l'adoption et sur la bioéthique. [...]
Pourtant, la perspective de l'adoption d'enfants par les homosexuels est bel et bien à l'horizon du projet du PACS.

b. **C. B.**: Adoption is legal for married couples as well as for single people beyond the age of twenty-eight [...] the adoption of a child by a homosexual couple, although it is not prohibited by the law, is limited by a ruling of the Conseil d'Etat from October 9 1996, which refuses to "give the right to adopt to a homosexual" [...] resorting to artificial insemination or *PMA* is the exclusive right of married couples and partners living together for two years. Regarding surrogacy, it is obviously prohibited by the law [...]. Officially, proponents of PACS refuse to leave to it the possibility of influencing people's status and state that these matters should be left for future laws on adoption and bioethics. [...]
Yet the horizon of the PACS project is in fact the adoption of children by homosexuals.

In fact, looking at Google Ngram Viewer for French and comparing the curves of "GPA" and "Gestation pour autrui", we can infer that the acronym has been in use for a long time before the phrase "gestation pour autrui". Adding the curve for "mère porteuse",

17 In case one doubts this, we can refer to Genesis 16:1-2: "Now Sarai, Abram's wife, had borne him no children. And she had an Egyptian maidservant whose name was Hagar. So Sarai said to Abram, 'See now, the Lord has restrained me from bearing children. Please, go in to my maid; perhaps I shall obtain children by her.' And Abram heeded the voice of Sarai." (*Holy Bible: The New King James Version* 1975)

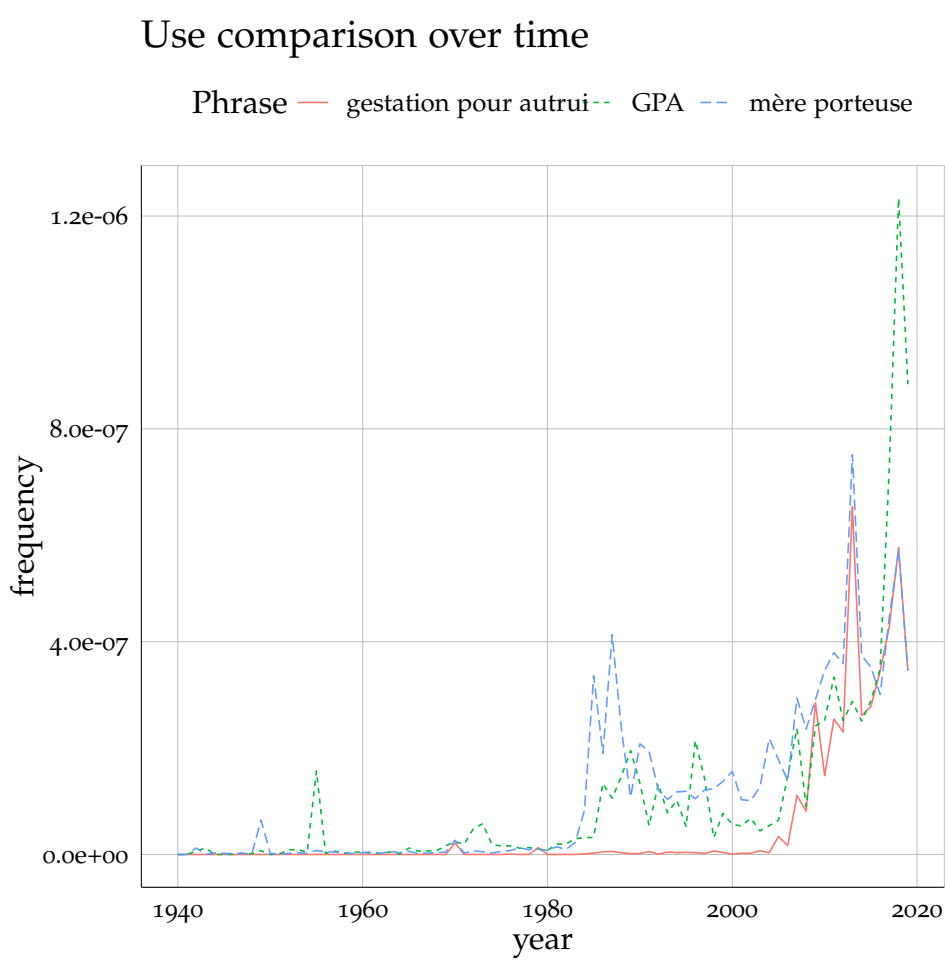


Figure 21: Comparison of phrases' usage using Google Ngram Viewer.

we can see that it has been much more widely used overall (Figure 21[18]). Although the "mère porteuse" curve does follow the "GPA" curve to some extent, if we look at the document results corresponding to the first peak in the use of "GPA", in the 1980s, we can see that most of those documents are works of engineering.

- The association of the acronym to the signification of "gestation pour autrui" seems to truly emerge during the 2000s, but it is likely that it had not entered public discourse until the 2013 debates on same-sex marriage (the MPT corpus); as a reminder, our pre-trained embeddings were trained on a web-crawled corpus, FrWac, which only contains pages that came before 2012. We can infer from this several things:

    1. In the general discourse, "GPA" had no particular connotations/meaning, its distribution being circumscribed to technical writings and the engineering sociolect. Starting in the 2000s it seems to be readily used at the same frequency as "gestation pour autrui", it has since largely exceeded both "gestation pour autrui" and "mère porteuse" in frequency of use, with a very important peak leading up to the year

18 With smoothing = 0 and restricted to the years 1940-2019. The original can be be viewed here: https://books.google.com/ngrams/graph?content=gestation+pour+autrui%2CGPA%2Cm%C3%A8re+porteuse&year_start=1940&year_end=2019&corpus=30&smoothing=0#

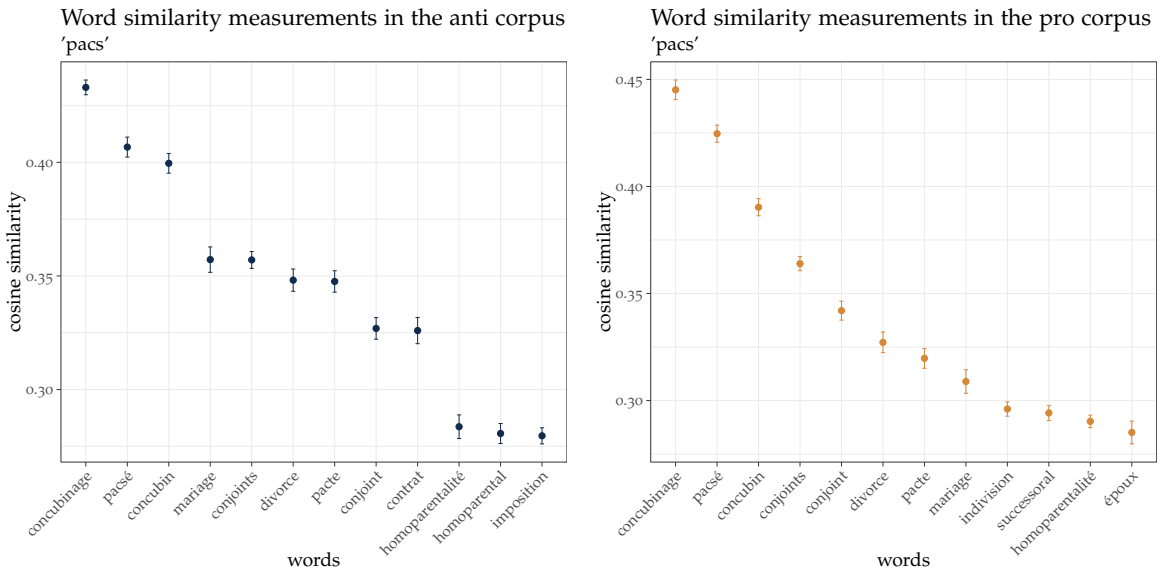


Figure 22: Comparison of closest semantic neighbours for *PACS* across MPT models. Overall layout identical to that of Figure 13.

2019, when the extension of IVF to lesbian couples was discussed in a new bioethics law at the Assemblée Nationale.

2. If we were to take our data at face value (which is very risky), we could say that diachronically, "GPA" has acquired its contemporary meaning over the years, with no particular associations during the PACS debates, but with a clearly defined meaning by the time of the MPT debates.

This last conclusion is however a bit strong given the overall results of the corpus.

Regarding the acronym "PMA", it has very few occurrences in the PACS corpus, which can be seen with the most similar words being extremely alike across corpora. Again, there is a possibility that this acronym has earned a lot of its connotations and meaning more recently, with greater use. A more interesting word to look at is "mariage": whereas "PACS" is among the most similar words to "mariage" for both the *pro* and *anti* in the MPT corpus, it only is for the *anti* in the PACS corpus. Let's look at how "PACS" itself looks in both corpora. The most similar words to "PACS" in the MPT corpora are shown in Figure 22, those for the PACS corpora are in Figure 23.

What we can see is that the word "mariage" is a closer nearest neighbour to "PACS" than "PACS" is of "mariage" for the *pro* in the PACS corpus, this difference is not as pronounced in the MPT corpus. In a few words, this means that the *pro* use the word "PACS" in a way that is very reminiscent of "mariage", but use "mariage" with a distribution that is in fact more similar to many other words. In a way, that sounds like a dogwhistley use of language, as it did for the "GPA"/"PMA" distinction in the MPT corpus: one word has different meanings (in a distributional sense) according to the community that uses it. In the *anti* community, "PACS" and "mariage" are semantically close, whereas in the *pro* community, they are used in distinct ways, to mean distinct things.

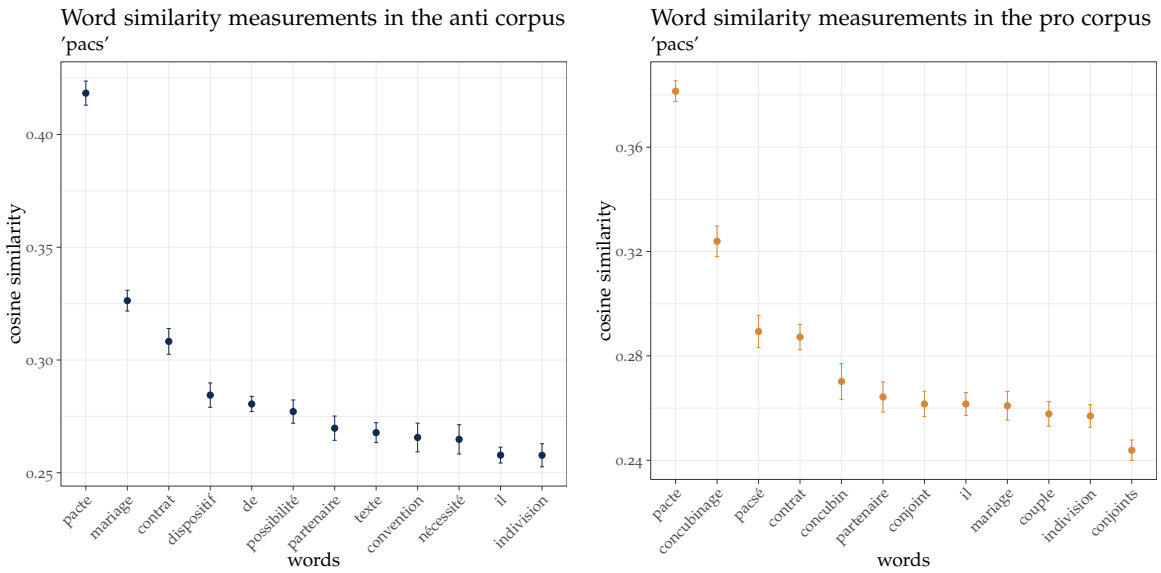


Figure 23: Comparison of closest semantic neighbours for *PACS* across PACS models. Overall layout identical to that of Figure 13.

The rest of the data related to the PACS corpus in presented in Appendix D.

8.3.3 *Measuring the degree of difference*

All of these inferences that we might want to make on the data that we observe are however irrelevant without a proper measure that allows us to say that these differences indeed exists, and quantifiably so. No matter how interesting the story-telling around it can be, none of this is of value without such a metric. As stated earlier, what we need is a way to measure not *how different* the lists of most similar words are, but *how interestingly different* they are. Although we have not managed to find such a metric, we have explored the topic, and the following is a starting point:

Given a word of interest w, we can generate one list of most similar words per sub-corpus, L_{pro} and L_{anti}. These lists are likely to be different in their orderings. Limiting ourselves to the top 12 most similar words, we can take the words that appear in L_{pro} but do not appear in L_{anti} and check their similarity *rankings* in the *anti* corpus.

If the measure works as planned, we should observe that in average these rankings are lower for our control words and higher for our test words.

The measure is asymmetrical, meaning that applying it to $\langle L_{pro}, L_{anti} \rangle$ will not give the same results as applying it to $\langle L_{anti}, L_{pro} \rangle$. We can therefore have three outcomes:

1. The 'missing words' from L_{pro} and L_{anti} both present similar rankings, these rankings are in average low and not very spread-out: this means that the words are used very similarly in both corpora.

2. The 'missing words' from L_{pro} and L_{anti} both present similar rankings, these rankings are in average high and not very spread-out: this means that the words are used very

dissimilarly in both corpora. Note: If the rankings are in fact very spread out, it might be an indication that a few of the 'missing words' are out of place, we will categorize these as belonging to the second case.

3. The 'missing words' from L_{pro} and L_{anti} present dissimilar rankings, these rankings are in average low for one and high for the other. This means that the set of uses of one word across corpora differs; specifically, the uses of one word in one corpus is a strict subset of the uses of that word in the other corpus. This could indicate that something dogwhistley is going on.

There is however another issue here. As per Figure 12, the list of most similar words is the *average* list of most similar words, meaning that we have computed the cosine similarities for all word-pairs, averaged them, and took the ranking of averaged cosine similarities. It might very well be the case that in some of the models we have generated, the rankings of a specific word are very different, or even worse, that the rankings of words differs a lot across models in general. If that is the case, then simply using the average cosine similarity measure to do the ranking appears to be unsatisfactory, especially in the case of a differentiation measure.

There are ways, however, to mitigate the feeling of unsatisfactoriness and take into account the possible variation in word ranking that results from our many models: for every "missing word", instead of taking its rank directly, we compute the 95% interval around its associated cosine similarity. We then take the highest average cosine similarity that falls within this 95% CI. The measurements we end up with should be much more conservative, but would also feel more meaningful, as they take into account the variation intrinsic to the measures we are making.

The description of the measure itself is complex enough, a step-by-step implementation using R is available at https://github.com/LangdP/difference-measure-attempt, with a lot more detail.

Let's call that measure M. Here are the results of M when applied in the MPT corpus, first on the word "amendement" (M_a), then on "mariage" (M_m):

$$M_a(L^a_{anti}, L^a_{pro}) = 23.28571$$
$$M_a(L^a_{pro}, L^a_{anti}) = 103.5714$$

$$M_m(L^m_{anti}, L^m_{pro}) = 23.85714$$
$$M_m(L^m_{pro}, L^m_{anti}) = 35.71429$$

As we can see, the results are not like what we expected. In fact it seems to be the case that the control word "amendement" is used in a dogwhistley way, which makes little sense. In any case, however, this measure gives us an idea of the kind of metric we would need in order to make sense of the data we have generated.

8.4 CONCLUDING REMARKS

We have attempted to use the insights from Chapter 7 and implement them directly onto two corpora that sounded like promising sources of dogwhistles. The results we have observed, although not uninteresting, are disappointing, and we are far from having a tool that allows us to quantify the "dogwhistleness" of a given term in a corpus. The experience has brought with it some interesting considerations, however. Specifically, we have underlined how techniques like *word embeddings* can be used to explore data and give us information in a way that makes them complementary with more common corpus analysis techniques. So far, the attempt has not proven to be very successful, but we notably reached a notion of the kind of measure that we would want to be able to successfully implement to assess the dogwhistleness of language, namely something like what is presented in Section 8.3.3.

One issue that remains is: why did it not work? There are several reasons that we can invoke to explain this:

1. One possibility is that we chose the corpora poorly. Although these corpora present interesting characteristics for the study of dogwhistles, it is very possible that there are in fact very few, or even no instances of dogwhistley language there. This is possible, but the accusation of pandering to, e. g., religious affects is made explicit in the corpus (see (34)), and some of the people involved in the debate were usually found at the forefront of the corresponding protests, walking along (very) openly Christian figures[19].

(34) a. **Henri Guaino** (Against *mariage pour tous*) : Vous ne voulez pas seulement que l'homme domine la nature. Vous voulez que le social triomphe de la nature et que sa victoire soit sans partage. Vous tournez ainsi le dos à la raison, car c'est la déraison qui commande à l'homme de vouloir nier sa nature. Où cela nous mènerait-il, sinon sur les voies les plus dangereuses ?
Henri Emmanuelli (Pro *mariage pour tous*) : On n'est pas à Rome !

b. **H.G.**: You don't merely want man to dominate nature. You want the social to vanquish nature and you want its victory to be complete. In so doing, you are turning

19 See for example this article, which underlines the presence at the same marches of representative Hervé Mariton (who does not actually identify as Catholic, but as Jewish) and fundamentalist Catholic activist Alain Escada and the Civitas organization: https://www.lemonde.fr/societe/article/2012/11/17/premiere-mobilisation-nationale-contre-le-mariage-gay_1792198_3224.html

> your back against reason, because it is madness which commands man to deny his nature. Where would that take us, except on the most dangerous paths?
> **H.E.**: This isn't Rome!

2. Another possibility is related to the data: maybe there simply isn't enough of it. While this is a possibility from the technical point of view, it is a very unsatisfactory response due to the fact that if we want to explore the possibility of dogwhistles in this corpus, there simply is no more data. Saying "there's not enough data" does not answer the problem, the only interpretation we can have of this is rather: WE and related techniques are not appropriate for that level of data, which means that they are not appropriate for many tasks in the humanities. The issue is not with the data, it is with finding other tools, more appropriate to answering these questions, but which also provide us with a similarly interesting notion of semantic distance.

3. Yet another possibility is that the issue is more on the technical side of things. It is possible that the pre-trained vectors were of poorer quality than anticipated, that we would need a bigger and better starting model, specifically one that has been trained up until, if not even beyond, the date of the corpora we are analyzing. Similarly, it is possible that the choice of hyperparameters that we made when training the models was not right, and perhaps we could optimize those, but this does not come without problems: first of all, the literature on this kind of work is not very large, and there is barely any discussion of proper hyperparameter tuning there, meaning that there is no standard, and most of the work in that area would be guesswork; second, optimize according to what? The use that we want to make of WE is very far from what is usually done with them. Specifically, it is hard to translate into a *task* with a *success rate*. As long as we do not have a proper performance metric that goes beyond mere human evaluation, then there is nothing to optimize towards. Besides, this is again against the principles discussed in Chapter 7, whereby we explicitly stated that this was to be seen as a way to *explore corpora*, and not to be thought as a task, at least for now.

4. In any case, and as one might have seen, there have indeed been technical issues in the overall process, notably regarding the tokenization and lemmatization. Our belief is that while more easy to solve, these issues probably do not have as big an influence on the end results as the others.

Overall, the attempt can be classified as a relative failure, but it can also be seen as a starting point for future work in the domain. While it is not the most satisfactory explanation to the limitations that we have observed in the results and their interpretation, the issue lying in the corpora themselves seems to us to be the most likely. If that happens to be the case and the corpus size simply does not allow one to use WE techniques for data exploration, then

other tools will have to be found for discourse analysis in general, as working on much larger corpora is likely to not be a possibility, at least if one is interested in the study of Parliamentary debates.

Part V

DISCUSSION AND CONCLUSION

9

DISCUSSION AND GENERAL CONCLUSION

"I always had hopes of being a big star. But as you get older, you aim a little lower. Everybody wants to make an impression, some mark upon the world. Then you think, you've made a mark on the world if you just get through it, and a few people remember your name. Then you've left a mark. You don't have to bend the whole world. I think it's better to just enjoy it. Pay your dues, and just enjoy it. If you shoot a arrow and it goes real high, hooray for you."

— Dorian Corey, *Paris is Burning*, most likely reflecting on how a good dissertation is a done dissertation. (Livingston, 1991)

"VLADIMIR:

Well, shall we go?

ESTRAGON:

Yes, let's go."

They stay.[1]

— Vladimir and Estragon in *Waiting for Godot*, illustrating how you're never really ready to leave the dissertation aside and move on with your life. (Beckett, 1952)

9.1 MAIN FINDINGS

Dogwhistles are an interesting phenomenon for pragmatics, and the literature about them is only getting richer and richer. This work has allowed us to explore a number of things about them.

First of all, we have managed to present them in known terms. The concept of dogwhistles is not primarily a linguistics concept, and an obvious challenge of this dissertation was trying to make it one. This has meant using the vocabulary of linguistics in order to characterize the dogwhistle communication phenomenon. Relying on insights from linguistics itself, we have managed to fuel our approach to dogwhistles using concepts from (formal) pragmatics (Chapter 2) and sociolinguistics (Chapter 3). These building blocks allowed us to shed some light on the existing literature on dogwhistles (Chapter 4).

Once dogwhistles were explained in terms that characterized other linguistic phenomena, we were able to give them a formal definition that effectively relied on existing work in formal

1 "VLADIMIR : Alors, on y va ? ESTRAGON : Allons-y. *Ils ne bougent pas.*"

linguistics and cognitive science. Using the tools from formal linguistics, we managed to give an account in mathematical terms of dogwhistle communication (Chapter 5), effectively reducing dogwhistles to special cases of other objects. This is one part of the formalization process: explaining phenomena that we do not understand using objects that we do understand.

Once these mathematical objects were defined, we observed that we could use them to characterize objects that were dogwhistle-*like*, as well as special cases of dogwhistles (Chapter 6), leading to the second important part of the formalization process: abstracting away from the phenomena under study and link them to other phenomena, effectively building a general object that can account for several different phenomena.

These considerations about the possible definition of dogwhistles have led us to a new understanding of them. That new understanding in terms of partial synonymy led to hypotheses regarding possible computational answers to the issue of discovering dogwhistles in a text (Chapter 7). Using the language of computational linguistics, we managed to see dogwhistles under yet another light, one that allowed us to focus no longer on the task of their description, but on that of their detection (Chapter 8). In this endeavor, we have not only proposed a way to automatically detect dogwhistles in a text, but also a new way of envisioning tools of computational linguistics, not merely as oriented on specific applied tasks, but as useful for giving us more fundamental insights in the definition of linguistic phenomena and explore linguistic data, rather than simply exploit it.

9.2 DISCUSSION

Even limiting ourselves to giving an understanding of dogwhistles in terms of formal pragmatics, it feels like this work has barely scratched the surface of the topic. In and outside of pragmatics, many questions remain to be answered, or even tackled. In this last leg of the thesis, we will attempt to briefly discuss some of these issues, including some thoughts that did not make it into the final dissertation as proper chapters, usually for lack of time. Mainly, the questions we could still ask are:

1. What is the strategy behind using dogwhistles?
2. Who benefits from using dogwhistles?
3. How do dogwhistles come into being?
4. Can any linguistic expression become a dogwhistle?
5. Do they disappear, how, why?

Some of these issues have been discussed in some way or other in the dissertation, some have not, it goes without saying that these supplementary remarks will not suffice to definitely answer any of these questions. A lot of what follows goes beyond the scope of proper research

and constitutes personal opinions and thoughts; it has been fueled by those last few years of discussing and reading about dogwhistles and political discourses in general.

9.2.1 *Words on the lifecycle of dogwhistles*

The content of this sections relies on tweets. To illustrate the discussion, the content of these tweets is presented as screenshots. Unless the person tweeting is a public figure or an organization, the tweets have been anonymized.

The work presented here, especially in Chapters 4, 5 and 6, proposed an analysis of a phenomenon whose existence was assumed. Of course, there is a lot of literature, dating back from a few decades now, that discuss dogwhistles, and attest their existence in political discourse, but the issue of how dogwhistles come to be and emerge is rarely discussed, and a mechanism for their coming to be is usually not mentioned. In their *function*, they present an obvious interest for actors in the political realm, but how does a specific expression become a dogwhistle? What qualifications does it need? Can any word become a dogwhistle[2]?

One thing can be said already given what we have seen in Chapters 5 and 6: dogwhistles are words that have differing meanings according to communities; supposedly any word that fills this description could become a dogwhistle and be used as such. This state of things can however be understood as either a pre-requisite or a resulting state of the use as a dogwhistle and gives us little indication as to what it looks like when it emerges. There are few works on this issue, but we can for example cite Åkerlund (2021) for the case of the term "kulturberikare" ("culture enricher") in Swedish politics.

9.2.1.1 *Birth*

Intuitively we can think of two ways that dogwhistles emerge in discourse. One is the repurposing of existing vocabulary. This is what we have focused on here and includes the case of "inner cities" or "wonder-working powers", these terms pre-exist their use as a political tool. In that case it can be that the meaning distinction on which our approach relies is in fact a pre-requisite to the dogwhistle formation. That dissimilarity can result from different prejudices in given communities, or different experiences of the world leading to a different labeling of things of the world. Another solution is that words are created that specifically refer to different things according to the community in which they are used. In a way, this is what "culture enricher" in the Swedish context is: the expression was first used by people in favor of immigration policies in a non ironic way, seeing an influx of foreign people and culture as something that would in the long term change Swedish culture in a positive way, towards more diversity; afterwards it was taken up by neo-nazi groups and the far right in

2 Note that this work has focused on noun phrases, but there is no reason to believe that dogwhistles are limited to any such syntactic category, especially when we consider that they do not even have to be limited to linguistic units.

an ironic way, referring to the metrics claiming that immigrant populations are more likely to commit crimes, particularly sex crimes (Åkerlund, 2021). Shortly after its inception, the term has come to mean different things for different groups.

Let's take a quick look at three examples from a French context.

'ENSAUVAGEMENT' "Ensauvagement", which could be translated literally as "wilding", "the process by which something gets more savage", is a word of the French language derived from *sauvage*, a term that can be thought of as ressembling the English "wild, brutal, rude, close to a state of nature", and has for some time been used to refer to foreigners. The term *ensauvagement*, although it has noticeably been used by Aimé Césaire to talk about colonizers[3], is now readily associated with the far right, and especially figures like Marine Le Pen. More specifically, the word has been associated with the far right thinker Laurent Obertone in his essay *La France Orange Mécanique*, published in 2013, but has since been taken up by several people at the Front National, Marine Le Pen's party[4]. Users of the term claim it has no racial undertones and is simply a statement about rising crime rates in France, but the history of the term as well as its appropriation by the far right seems to indicate otherwise. A cover from the magazine Valeurs Actuelles (Figure 24) illustrates this well.

How is it dogwhistley? It has acquired dogwhistle hues in 2020 when it was used in earnest by Minister of the Interior Gérald Darmanin, a figure usually associated with the more Conservative right wing, but not necessarily with the far right. Since then it has re-entered the public discourse and is used to openly refer to the idea of rising crime rates, but is thought to also have strong racial undertones. This has led to the word becoming a fairly popular hashtag, still in use as of the time of writing (24/12/2021[5]).

In this scenario, a word that exists in the language was taken up by the far right and has become a term that is readily associated with their manner of speech; it has since been taken up by more moderate conservative figures, and is now used more largely in the public discourse. We are in the case of a word acquiring some dimension of dogwhistleness.

We could ponder about the strategy behind the use of that word by Darmanin. One possibility is that the word was used with a genuine reference to rising crime rates and no intention to pander to the far right. Of course this is what a standard use of dogwhistles would aim towards. Given the known history of the word, its etymology and its rather recent uses by the far right, this seems unlikely. Another solution could be that Darmanin seeks to gather the support of the far right without alienating other voters, i. e., he is dogwhistling. Given that the

3 "…at the end of the galvanized racial pride, of the pervasive crowing, there is the poison instilled in the veins of Europe and the slow but steady progress of the *ensauvagement* of the continent." "…au bout de cet orgueil racial encouragé, de cette jactance étalée, il y a le poison instillé dans les veines de l'Europe et le progrès lent, mais sûr, de l'*ensauvagement* du continent", in Césaire (1950/1955), emphases mine.

4 This article from Le Monde retraces the history of the word in contemporary politics: https://www.lemonde.fr/les-decodeurs/article/2020/09/03/l-ensauvagement-un-mot-a-l-histoire-sinueuse-surtout-utilise-par-l-extreme-droite_6050851_4355770.html.

5 Happy holidays to you too if you happen to read this on December 24th.

Figure 24: The tweet presenting the front page for an issue of Valeurs Actuelles, discussed in footnote **??**.

Front National nowadays has a very strong voter base, it is not unlikely that the governing party should seek the support of that part of the population by using that kind of language. But something does not add up: this message is not exactly "hidden". Its ties with the far right are known and its use by Darmanin has been commented upon directly after its use. We can therefore see it in another way: Darmanin is either attempting the "Punch, Kick, Parry" strategy mentioned in section 4.1.1, seeking to cause reaction in the opposition; or it could be that in fact the use of the term has become more "mainstream", even acceptable, which might be a long-term effect of sugarcoated vocabulary and hedging including by far right parties themselves (see section 9.2.2 for an attempt at a discussion on this).

'GILETJAUNISATION' The word "giletjaunisation" is a word that has only recently appeared in the public discourse, as a reference to the Gilets Jaunes grassroots social movement in France, which was born in October 2018. The word can be taken to literally mean "acquiring properties related to the Gilets Jaunes movement". What these properties are exactly depends on who uses the word. Supporters of the movement have been seen as equating it to a general feeling of large-scale grassroots uprising (Figure 25), but other uses have been commented upon as being either somewhat equivalent to "conspiracy theorist" (Figure 26) or associated with a more general contempt for lower classes (Figure 27). It is unclear so far whether "giletjaunisation" is or might become a dogwhistle to some degree, it is possible however that it ends up referring to both a political movement/ideology and to a general disdain regarding a social group/some degree of parisianism according to who says it. So far, a general discourse strategy built around it seems hard to imagine, but this time we are in a case of a word being created and then immediately seeing its meaning split across communities.

'LE FINANCIER' On August 29th 2021, presidential candidate Jean-Luc Mélenchon tweeted, referring to an utterance from a speech he was giving that day: "No, the ennemy is not the muslim, it's the banker!"[6] (Figure 28). At the time of utterance, the sentence was commented upon by the French Jewish community as being a possible anti-semitic dogwhistle. The lack of convincing response to the accusations from the Mélenchon team did not alleviate the worries. In this case, we have a situation where the issue of religious individuals is brought up in the same sentence, and then the figure of the banker, typically associated with Jewish people, is mentioned. This is reminiscent of the "inner cities" case, because there again, there is a mainstream meaning to 'banker' that has nothing to do with Jewish people. In addition to that, the term can be understood as denoting bankers in general, which are generally representative figures of capitalism and therefore intrinsically opposed to the likes of Mélenchon (who is labelled as a far left candidate); but it can also denote Emmanuel Macron himself, who is known to have worked as a banker before his career in politics.

6 "Non, l'ennemi ce n'est pas le musulman, c'est le financier !"

Figure 25: Tweets using "giletjaunisation" as synonymous for "grassroots social movement".

Figure 26: Tweet equating "giletjaunisation" to conspiracy theory thinking.

Figure 27: Tweet where "giletjaunisation" is analyzed as a classist term.

Figure 28: The Mélenchon tweet mentioned in Section 9.2.1.1.

This case is interesting in terms of dogwhistle studies because it is a good illustration of the difficulty of identifying dogwhistles. In the case of "inner cities", the strategy is well documented and has been observed several times, but in this case, it is not nearly as present in public discourse, and in any case, this specific example comes with strong plausible deniability. Looking at the American examples we have seen, it looks like the dogwhistling here is a plausible explanation, but it is likely that we will need to see how everything unfolds from now on and whether there are other occurrences of similar possible dogwhistling happening on the part of Mélenchon and his supporters to see whether we are right in thinking that. Proper identification takes time.

9.2.1.2 *Use and decay*

What happens with dogwhistles once they are discovered? The intuitive response to this is: they stop being dogwhistles. If dogwhistles are a way to communicate a hidden message to some people in the audience, then surely their being discovered means that they stop carrying a *secret* message. Several things can follow from this:

1. People who used them with the intent of secretly referring to ideas or ideologies deemed unacceptable in the public space stop using it, since they can no longer have that purpose. In case one wants to continue the dogwhistling, then one has to find new dogwhistles.

2. The dogwhistles keep being used, possibly in the context of the "Punch, Parry, Kick" strategy.

3. The dogwhistles lose their attachment to a specific community and become part of the standard public speech, either because the original double-meaning associated with a specific community disappears or because the norms of discourse have been redefined (Section 9.2.2).

An interesting older term that could be thought of as dogwhistley in the French context is "jeunes de banlieue", "youth from the suburbs". The expression has a compositional meaning that is clear enough, yet it is in fact never used in this compositional way. Its meaning ranges from "some youth from some suburbs" (specifically underpriviledged people from disadvantaged suburban areas) to "immigrant youth of African descent", due to the fact that the more disadvantaged suburban areas have historically been used for the housing of post-World War II waves of immigration; the point is that it just about never means "youth from Neuilly-sur-Seine[7]". See e. g., Castel (2006), Derville (1997), Guénolé (2015), and Longhi (2012) for some discussions on this figure of the French public discourse. The use of the word has known a peak in the year 2005, interestingly at the same time the word "ensauvagement" has known one, see Figure 29[8], probably due to what is now sometimes called the "émeutes de

7 A fancy suburb in the West of Paris.

8 Original here https://books.google.com/ngrams/graph?content=jeune+de+banlieue%2Cjeunes+de+banlieue%2Cjeunes+de+banlieues%2Cjeunes+des+banlieues&year_start=1940&year_end=2019&corpus=30&smoothing=0&

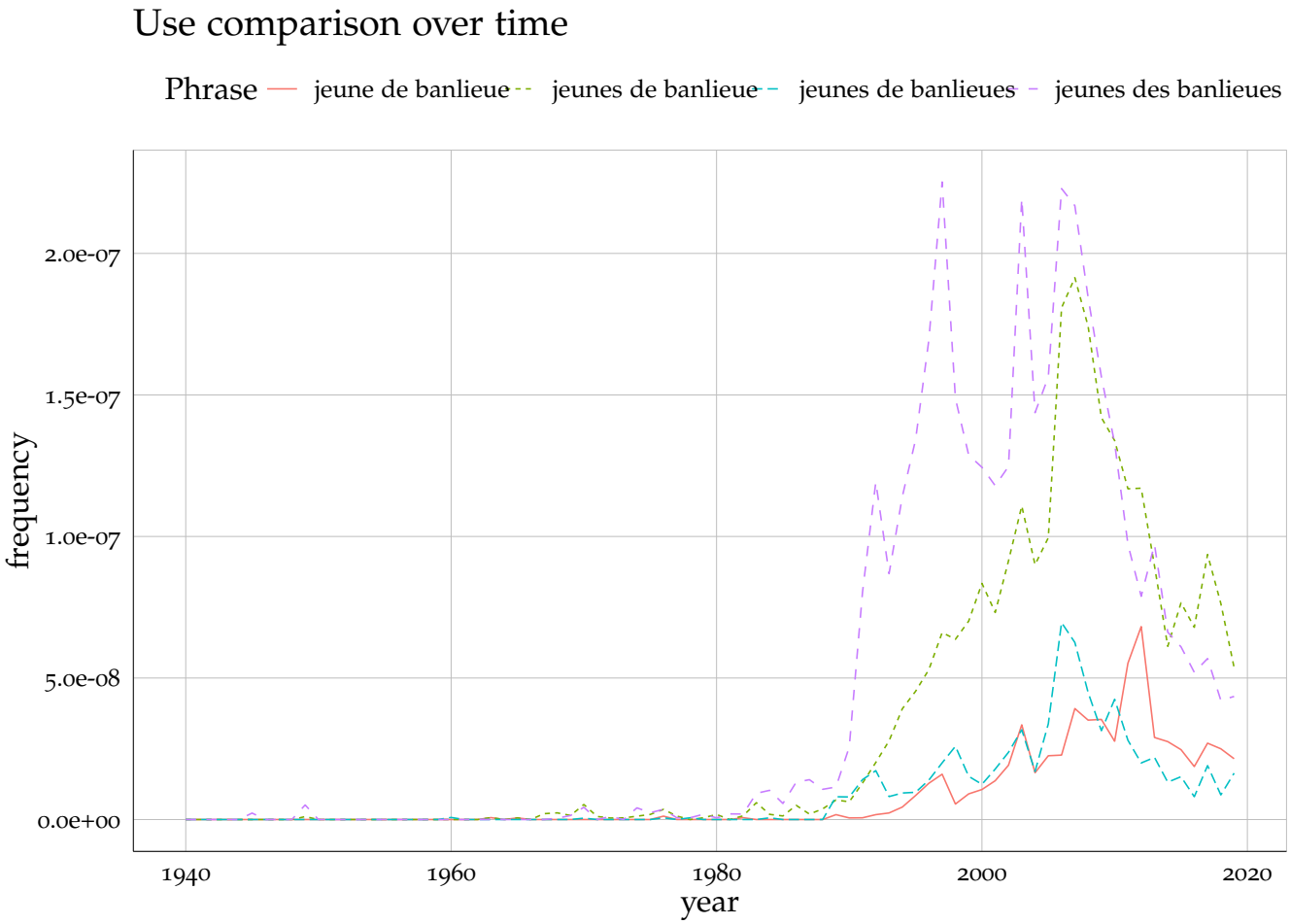


Figure 29: Comparison of phrases' usage using Google Ngram Viewer (*jeune* de banlieue**)

2005", "2005 riots", following a series of events including the deaths of Zyed Benna and Bouna Traoré, respectively 17 and 15 years old, when they attempted to flee from the police during a chase by hiding in an electrical transformer. Following their deaths, a series of popular uprisings took place in many of the suburbs of Paris, especially in Seine-Saint-Denis (which covers the northernmost suburban areas of Paris).

The noun phrase "jeunes de banlieue", when used by conservative-leaning politicians and far right parties, is likely to be associated with immigration, or more recently the notion of Great Replacement[9], whereas it is likely that when used by left-leaning politicians, it rather refers to underpriviledged youth. In any case, while this might have been a dogwhistle at some point (with hidden racial undertones), the meaning of this noun phrase is nowadays transparent. It has not disappeared, but few people are fooled by it and in fact, the expression's

`direct_url=t1%3B%2Cjeune%20de%20banlieue%3B%2Cc0%3B.t1%3B%2Cjeunes%20de%20banlieue%3B%2Cc0%3B.t1%3B%2Cjeunes%20de%20banlieues%3B%2Cc0%3B.t1%3B%2Cjeunes%20des%20banlieues%3B%2Cc0`

9 See for example the recent (at the time of writing) intervention by Eric Zemmour:
"You have seen for yourselves the analysis in France Stratégie from a few months ago. In the *quartiers* of Paris and other larger cities' suburbs, we have in fact reached the end of the Great Replacement. There are 80% of young people between the ages of 0 to 18 that descend from extra-European immigration."
"Vous avez vu l'analyse de France Stratégie d'il y a quelques mois. Dans les quartiers des banlieues parisiennes ou autres grandes métropoles, nous sommes arrivés en fait à la fin du grand remplacement. Il y a 80% des jeunes entre 0 et 18 ans qui sont issus de l'immigration extra-européenne" in https://www.youtube.com/watch?v=acGYPoTUrJU.
Note the use of *quartiers*, "neighborhoods", which has followed a similar trend, though its original meaning is still largely dominant.

compositional meaning is never invoked. The term is such an important part the French public discourse however, that one could consider the idea that the figure of the "jeune de banlieue" has become some sort of persona (in the social meaning sense presented in Chapter 3), with stereotypically associated clothing style, attitudes and speech patterns.

9.2.2 *Dogwhistle strategizing*

Seeing how complicated they are to describe, and inferring how complicated they can be to properly come up with and use, one may wonder: why dogwhistles? Why are they used when they seem so obviously inefficient?

In Chapter 4, we have briefly mentioned several ways of using dogwhistles. One of those ways was characterized by dogwhistles being a follow-up to whistle-stop campaigning (Goodin and Saward, 2005): people who run for any kind of office need to gather as much support as they can in the population, and will therefore need to convince as many people as possible that they will satisfy their demands. In the absence of mass media, this takes the form of the candidate changing their discourse locally to entice as many people as possible to vote for them at any point in their campaign. With the advent of mass media and the broadcasting of official discourse, this solution is made obsolete, a new way to satisfy as many people as needed emerges in the form of dogwhistles, with politicians doing their best to satisfy as many (possibly contradictory) demands at once. This is reminding of the case of protean dogwhistling that we discussed in section 6.3.2. In this case, it is theoretically in the speaker's interest to use as vague a vocabulary as possible, so that people will fill in the blanks themselves, hopefully with the explanation they favor the most. Ambiguity can be safer than clarity.

This is not unlike what is sometimes referred to as *stonewalling* or *doublespeak*, and while it might be a nice way to evade an unwanted question, it is a risky overall strategy if one is campaigning, as the message that the politician sends in the end is clear to no one. While this seems to be broadly advantageous strategically, it feels like it is an approach that quickly reaches its limits[10].

The seemingly rational thing to do in politics would be to be excessively ambiguous, always, to offend no one and ensure the support of as many people as possible; this is not what happens, some content is communicated, sometimes unambiguously. This is for a simple reason: while it is necessary to gather *as much* support as possible, it is however impossible (and therefore excluded) to gather the support of *everyone*. Let's put it another way; the goal of poltical campaigning is to convince:

1. People who are uncertain who they are going to vote for among the possible candidates.
2. People whose ideology is not represented by any of the possible candidates (and perhaps do not vote as a result).

10 Besides, dogwhistles are associated with a notion of coded or hidden message, which is not necessarily associated with stonewalling and doublespeak.

The crowd of people who do vote and have a clear idea about the identity of the person they are going to vote for does not need convincing, either because they are already convinced, or because they will not be convinced. Which group is the most numerically important probably depends on many factors, but my impression is that most if not all political campaigning is targeted at group 1: they are involved in the political life to the point of voting, but not enough to not be budged from their positions. There are many reasons for the second group to exist, that will vary a lot depending on where one is. Some of the people who do not vote do it because of personal and political convictions, and those are people who will likely not be convinced by any candidate; others do it for lack of representation in the political game. Those that do not find adequate representation in the available candidates are likely to have ideologies that are not represented by mainstream parties and be considered "fringe". Among those ideologies, some are not represented because they have been shunned from the public debate, for historical or social reasons. Because they have been shunned from the debate, a candidate wanting to secure the votes of group 1 should not discuss them; but to secure the votes of group 2, they should be alluded to. Hiding the message is necessary to appeal without alienating.

Before discussing any further, it is probably worth reminding the following: in this day and age, and very much so in two-party political systems, it is very tempting to divide the political landscape into "the extremes" and "the rest", meaning mainstream political movements. This is simplistic, and there are good reasons to think that it is wrong (and in fact is very much a position that plays in the interests of mainstream political movements). There are countless political communities and groups of interest, of countless different sizes, some focused on a very specific issue, others adhering to entire complex representations of the world. The point of view that opposes "extreme" parties to "mainstream" parties is very misleading, presenting the "mainstream" or "center" as some sort of ideally balanced position, a position that all discussions should eventually more or less converge towards, an unmovable beacon of reason that people should strive towards reaching. It is probably more interesting to see the political landscape not as a space of few dimensions in which there is someone occupying clearly the center spot, but as a space of very many dimensions where there is no clear center, but many movements scattered everywhere, with a few sometimes clustering together. Similarly, a simple left-right dichotomy, though it does refer to extensionally well-defined movements, maintains that illusion of a "neutral" center. The political "center" is hardly neutral; in fact it means very little for a political movement to be "neutral", politics is about taking sides and making decision, it never is about neutrality.

Similarly, the idea that there is an unalterable set of politically acceptable policies and issues that would in fact constitute the political mainstream, is questionable. Such a set, if its exists, certainly changes over time, possibly over short periods, and defining why a given idea is or is not acceptable is not an easy task. One might want to resort to the concept sometimes called Overton window here. In a few words, this concept is used to refer to the set of politically

acceptable policies and debatable issues. Note that this is a common understanding of the concept, but that it is not necessarily what was meant by Joseph Overton. In fact it is unclear what was meant by Overton himself, because there are no sources by him mentioning the concept, although it is largely attributed to him by the Mackinac foundation, the think tank that he belonged to[11]. When used by its initiators, the concept means something much more specific, with the idea that all public policies can be ranked along a specific axis ("more free" to "less free"), the "window" is the section of this axis that can be openly debated in public. The goal of lobbyists at the Mackinac foundation and other institutions is to move or expand that window. In this case, a less precise definition of the window actually makes it more useful. "Freedom" (of what? From what?) is not in fact the important point, the important point is that the window can move and expand.

The idea that the political mainstream changes with time is not revolutionary or new, the question is: who opens the window? According to lobbyists at the Mackinac foundation, the lobbyists do it, and politicians are merely followers who will tell the public whatever they want to hear in order to get elected. Once again, this is simplistic. It is likely that a lot of the discourse can participate in moving or opening the window, but it seems unlikely that politicians do not do it and merely follow. Politicians do have interests and ideologies, and all of them have an interest in moving the window towards the ideas defended by those ideologies.

In the case of dogwhistles, some communities that are not represented by political candidates that suit them are still closer to some part of the mainstream than to others, at least on some issues. This translates into the fact that not any mainstream politician can tap into the voting power of any isolated community. In this sense, we can see the use of dogwhistles as an attempt at a partnership between the politician and the isolated community over the idea that if they come into power, they will open the Overton window to include some of that community's key issues. This partnership has to be credible, however, which is why trying to appeal to Evangelicals worked for Bush, but not for Clinton (see Section 4.2). Dogwhistles are not a secret cheat code that allows one to rally any community, they are a carefully planned political strategy to expand one's electorate, but that expansion follows some degree of continuity.

And who really wins at this game? Again, mainstream political parties do not have unmoving, unalterable ideologies, and if you look long enough into the abyss of socially unacceptable ideologies, the abyss also looks into you. This leads us to the case of the Nick Griffin use of dogwhistles[12]. In that case, the dogwhistle strategy emanates *from* the ingroup and is aimed towards the outgroup. The hopes here are likely to be a bit different. In the previous case, the goal of using dogwhistles was gathering votes for a specific event (such as an election), but in that second case, it comes from parties that typically – and this was all the more true at the time of Nick Griffin at the BNP – do not win general elections. The goal might be mere

11 See https://www.mackinac.org/12481 and https://www.mackinac.org/overtonwindow.

12 We could find other personalities from the far right, both in Europe and the USA, of course, but the quote that we presented in Section 4.1 is a good entry point.

sugarcoating of ideas in the hopes of convincing some voters, but if we think of it in terms of long-term goals and policies, like Griffin does, the goal is to make the ideas from the party *relevant* and *acceptable*, in other words, to open the Overton window. In a way, this is similar to the way we commented upon Richard Nixon's racial policies in Chapter 4, and part of the so-called *Southern strategy*, whereby the use of hedging allowed to put forward policies that had the end-goal of affecting Black people more than White people.

This is in part what is alluded to in Saul (2017, 2019), where we can probably start to decipher a pattern. *Dogwhistles* are an entry point. Once they have become the norm, there is no need to hide. The USA has recently made the experience of this under the presidency of Donald Trump, where a large part of the discourse that would have been dogwhistled before then was instead fairly open, specifically concerning immigration. Once the dogwhistles are no longer needed, *figleaves* – openly problematic[13] comments whose problematic content is simply not acknowledged by the speaker in spite of largely being there – come into play. Called "dogscreams" in Filimon (2016), these are instances of discourse where the mere mention of them not being racist is supposedly sufficient to make them acceptable in the public discourse. In any case, those set a precedent and are a clear marker that the norms of discourse have indeed changed.

9.3 CONCLUDING THE CONCLUSION

As stated above, we have here barely scratched the surface of dogwhistles, even though it feels like we have done a lot. My hopes are however that this dissertation has given *some* understanding of *some* flavor of dogwhistles. We have presented an object that is highly uncommon from a linguistics standpoint (Chapter 4), a way to seemingly send several messages with a single signal, to appear to be both one thing and another to different people; we have attempted to describe this object in very precise terms and reduce it to simpler, more well-known processes, with some degree of success (Chapter 5); we have used these abstractions and extended them to link the concept to broader phenomena in natural language, phenomena that can be seen, to an extent, as one way of constructing personae in the discursive space (Chapter 6) and enrich existing accounts of the phenomenon of social signaling. We have even (though with questionable success) attempted to pave the way for new uses of existing tools for linguistic inquiry (Chapter 7, Chapter 8). And yet this feels like a mere introduction to the subject.

Beyond all of this, there are many things at play. The formal descriptions that we have put forward focused on a cognitive view of language use and understanding – as is often the case in formal linguistics – but in a way they obscure a very important aspect of dogwhistles, which is their use as part of more general phenomena of *propaganda*, and more general phenomena of

13 The work in Saul (2017) talks specifically about racist comments, but I see no reason why this could not be extended to all forms of socially shunned discussions.

using language to influence others. Dogwhistles are important because they are not merely a way to gather votes, they can be a way of changing the general content of political discussions. One should watch out for them, find them, deconstruct them. The general shape of political discussions is changed over time in small increments, dogwhistles participate in those small increments, and if one does not want to see the window be fully open in some direction, one should pay attention not merely to *what* is said, but to *what is said by who to whom*, and *for what purpose*. There can be no healthy political discussion unless this is known of as many people as is possible.

Regarding the field of linguistics and pragmatics, discussing the idea of the construction of social identity through discourse, and the possibility for this identity to be many-faceted leads us to more profound discussions on the social space in which language users evolve. Formal linguistics has brought interesting representations from a cognitive standpoint, but too many issues in this subfield have been left aside because they resist explanations in those terms. Identity is one such issue, which is barely starting to be discussed by the formal linguistics crowd, and brings along with it a host of different questions, often underlining the very real limitations of existing accounts of language in formal approaches. It is important to underline that *ideal* language, *private* language, are interesting abstractions, but are ultimately not an adequate rendition of how language works, evolves and is used. Language is its users. The users live in a world. The world and experiences of the world feed the language. There can be no adequate portrait of language, even reduced to a cognitive ability, without taking into account all of those things, and there is no solution of continuity between them.

If on the way there, one can uncover deception and manipulation, then surely this is worth doing.

BIBLIOGRAPHY

Acton, Eric (2017). "Pragmatics, the third wave and the social meaning of definites." In: *Proceedings of New Ways of Analyzing Variation (NWAV)* 46.

— (2019). "Pragmatics and the social life of the English definite article." In: *Language* 95(1), pp. 37–65. DOI: doi:10.1353/lan.2019.0010..

Albertson, Bethany L. (2015). "Dog-Whistle Politics: Multivocal Communication and Religious Appeals." In: *Political Behavior* 37. ISSN: 1573-6687. DOI: 10.1007/s11109-013-9265-x.

Anthony, Laurence. *AntConc (Version 3.5.9) [Computer Software]*. Tokyo, Japan. URL: https://www.laurenceanthony.net/software.

Antoniak, Maria and David Mimno (2018). "Evaluating the stability of embedding-based word similarities." In: *Transactions of the Association for Computational Linguistics* 6, pp. 107–119.

Asher, Nicholas (1993). *Reference to Abstract Objects in Discourse*. Studies in Linguistics and Philosophy. Springer Netherlands. ISBN: 9789401117159. URL: https://books.google.fr/books?id=Rp7-CAAAQBAJ.

Asher, Nicholas and Alex Lascarides (2003). *Logics of Conversation*. Cambridge University Press.

Asher, Nicholas and Soumya Paul (2018). "Strategic Conversations Under Imperfect Information: Epistemic Message Exchange Games." In: *Journal of Logic, Language and Information* 27.4, pp. 343–385. ISSN: 1572-9583. DOI: 10.1007/s10849-018-9271-9. URL: https://doi.org/10.1007/s10849-018-9271-9.

Asher, Nicholas, Soumya Paul, and Julie Hunter (2021). "Bias in semantic and discourse interpretation." In: *Linguistics and Philosophy*. ISSN: 1573-0549. DOI: 10.1007/s10988-021-09334-x. URL: https://doi.org/10.1007/s10988-021-09334-x.

Asher, Nicholas, Soumya Paul, and Antoine Venant (2017). "Message Exchange Games in Strategic Contexts." In: *Journal of Philosophical Logic* 46 (4), pp. 355–404. ISSN: 1573-0433. DOI: 10.1007/s10992-016-9402-1. URL: https://doi.org/10.1007/s10992-016-9402-1.

Asher, Nicholas and Laure Vieu (2005). "Subordinating and coordinating discourse relations." In: *Lingua* 115.4. Coordination: Syntax, Semantics and Pragmatics, pp. 591–610. ISSN: 0024-3841. DOI: https://doi.org/10.1016/j.lingua.2003.09.017. URL: https://www.sciencedirect.com/science/article/pii/S0024384103001475.

Athenot, Eric (2009). "Walt Whitman, passant moderne." In: *Caliban. French Journal of English Studies* 25, pp. 139–152.

Austin, John Langshaw (1962). *How to Do Things with Words*. Clarendon Press.

Bach, Kent (Aug. 1999). "The Myth of Conventional Implicature." In: *Linguistics and Philosophy* 22, pp. 327–366. DOI: 10.1023/A:1005466020243.

Bach, Kent (Jan. 2008). "Pragmatics and the Philosophy of Language." In: *Handbook of Pragmatics*. Ed. by Laurence Horn and Gregory Ward. John Wiley and Sons, pp. 463 –487. ISBN: 9780470756959. DOI: 10.1002/9780470756959.ch21.

Bachmann, Ingo (2011). "Civil partnership–"gay marriage in all but name": A corpus-driven analysis of discourses of same-sex relationships in the UK parliament." In: *Corpora* 6.1, pp. 77–105.

Baker, Paul (2004). "'Unnatural Acts': Discourses of homosexuality within the House of Lords debates on gay male law reform." In: *Journal of sociolinguistics* 8.1, pp. 88–106.

— (2008). *Sexed texts: language, gender and sexuality.* Equinox.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta (2009). "The WaCky wide web: a collection of very large linguistically processed web-crawled corpora." In: *Language resources and evaluation* 43.3, pp. 209–226.

Beaver, David and Jason Stanley (2018). "Toward a Non-Ideal Philosophy of Language." In: *Graduate Faculty Philosophy Journal* 39 (2), pp. 503–547. DOI: https://doi.org/10.5840/gfpj201839224.

Beckett, Samuel (1952). *En attendant Godot*. Les Éditions de Minuit.

Beltrama, Andrea (2018). "Precision and speaker qualities. The social meaning of pragmatic detail." In: *Linguistics Vanguard* 4.1, p. 20180003. DOI: doi:10.1515/lingvan-2018-0003. URL: https://doi.org/10.1515/lingvan-2018-0003.

Beltrama, Andrea and Laura Staum Casasanto (2017). "Totally tall sounds totally younger: Intensification at the socio-semantics interface." In: *Journal of Sociolinguistics* 21.2, pp. 154–182. DOI: https://doi.org/10.1111/josl.12230. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/josl.12230. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/josl.12230.

Bender, Emily M. and Batya Friedman (Dec. 2018). "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science." In: *Transactions of the Association for Computational Linguistics* 6, pp. 587–604. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00041. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00041/1567666/tacl_a_00041.pdf. URL: https://doi.org/10.1162/tacl_a_00041.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922. URL: https://doi.org/10.1145/3442188.3445922.

Benz, Anton, Gerhard Jäger, and Robert van Rooij (Jan. 2006). "An Introduction to Game Theory for Linguists." In: pp. 1–82. ISBN: 978-1-349-52317-7. DOI: 10.1057/9780230285897_1.

Béraud, C., P. Portier, P. Guyot-Sionnest, M. Wieviorka, and J. Ténédos (2015). *Métamorphoses catholiques: Acteurs, enjeux et mobilisations depuis le mariage pour tous*. Éditions de la Maison des sciences de l'homme, Paris. ISBN: 9782735126965. URL: https://books.google.fr/books?id=5tnxDwAAQBAJ.

Bergen, Leon, Roger Levy, and Noah Goodman (2016). "Pragmatic reasoning through semantic inference." In: *Semantics and Pragmatics* 9. ISSN: 1937-8912. DOI: http://dx.doi.org/10.3765/sp.9.20.

Bhat, Prashanth and Ofra Klein (2020). "Covert Hate Speech: White Nationalists and Dog Whistle Communication on Twitter." In: *Twitter, the Public Sphere, and the Chaos of Online Deliberation*. Ed. by Gwen Bouvier and Judith E. Rosenbaum. Cham: Springer International Publishing, pp. 151–172. ISBN: 978-3-030-41421-4. DOI: 10.1007/978-3-030-41421-4_7. URL: https://doi.org/10.1007/978-3-030-41421-4_7.

Bieber, Justin, Jason Boyd, and Mason Levy (2015). *What Do You Mean?*

Boleda, Gemma (2020). "Distributional Semantics and Linguistic Theory." In: *Annual Review of Linguistics* 6.1, pp. 213–234. DOI: 10.1146/annurev-linguistics-011619-030303. eprint: https://doi.org/10.1146/annurev-linguistics-011619-030303. URL: https://doi.org/10.1146/annurev-linguistics-011619-030303.

Boltanski, Luc and Laurent Thévenot (1987). *De la Justification : les Économies de la Grandeur*. Paris: Gallimard.

Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai (2016). "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." In: *Advances in neural information processing systems*, pp. 4349–4357.

Bonikowski, Bart and Yueran Zhang (2020). "Populism as Dog-Whistle Politics: Anti-Elite Discourse and Sentiments toward Out-Groups." In: *SocArXiv*. DOI: https://doi.org/10.31235/osf.io/m29kf.

Buffon, G.-L. Leclerc comte de (Aug. 1753). *Discours sur le Style*. Discours prononcé à l'Académie Française par Monsieur de Buffon, le jour de sa réception, le 25 Août 1753. URL: https://fr.wikisource.org/wiki/Discours_sur_le_style.

Burnett, Heather (2017). "Sociolinguistic interaction and identity construction: The view from game-theoretic pragmatics." In: *Journal of Sociolinguistics* 21.2, pp. 238–271. DOI: https://doi.org/10.1111/josl.12229. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/josl.12229. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/josl.12229.

— (Oct. 2019). "Signalling games, sociolinguistic variation and the construction of style." In: *Linguistics and Philosophy* 42.5, pp. 419–450. ISSN: 1573-0549. DOI: 10.1007/s10988-018-9254-y. URL: https://doi.org/10.1007/s10988-018-9254-y.

— (in press). *Meaning, Identity and Interaction: Sociolinguistic Variation and Change in Game-Theoretic Pragmatics*. Cambridge University Press.

Bush, George W. (Jan. 28, 2003). *Address Before a Joint Session of the Congress on the State of the Union*. State of the Union address pronounced by George W. Bush on the second year of his first presidency. URL: https://www.presidency.ucsb.edu/node/211931 (visited on 06/30/2021).

Caldwell, Thomas Price (2018). *Discourse, Structure and Linguistic Choice*. Springer International Publishing. ISBN: 978-3-030-09231-3. DOI: 10.1007/978-3-319-75441-3.

Calfano, Brian Robert and Paul A. Djupe (2009). "God Talk: Religious Cues and Electoral Support." In: *Political Research Quarterly* 62.2, pp. 329–339. DOI: 10.1177/1065912908319605. eprint: https://doi.org/10.1177/1065912908319605. URL: https://doi.org/10.1177/1065912908319605.

Cameron, Deborah (2001). *Working with spoken discourse*. Sage.

Campbell-Kibler, Kathryn (Feb. 2007). "ACCENT, (ING), AND THE SOCIAL LOGIC OF LISTENER PERCEPTIONS." In: *American Speech* 82.1, pp. 32–64. ISSN: 0003-1283. DOI: 10.1215/00031283-2007-002. eprint: https://read.dukeupress.edu/american-speech/article-pdf/82/1/32/395097/ASp82.1.2Campbell-Kibler.pdf. URL: https://doi.org/10.1215/00031283-2007-002.

— (2008). "I'll be the judge of that: Diversity in social perceptions of (ING)." In: *Language in Society* 37.5, 637–659. DOI: 10.1017/S0047404508080974.

Carroll, Lewis (1872). *Through the Looking Glass and What Alice Found There*. London: Macmillan and Co. ISBN: 1546659196. URL: https://en.wikisource.org/wiki/Through_the_Looking-Glass,_and_What_Alice_Found_There.

Castel, Robert (2006). "La discrimination négative. Le déficit de citoyenneté des jeunes de banlieue." In: *Annales. Histoire, Sciences Sociales* 61.4, 777–808. DOI: 10.1017/S0395264900030407.

Champagne, John (2008). "Walt Whitman, our great gay poet?" In: *Journal of homosexuality* 55.4, pp. 648–664.

Clark, Stephen (2015). "Vector Space Models of Lexical Meaning." In: *The Handbook of Contemporary Semantic Theory*. John Wiley and Sons, Ltd. Chap. 16, pp. 493–522. ISBN: 9781118882139. DOI: https://doi.org/10.1002/9781118882139.ch16. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118882139.ch16. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118882139.ch16.

Combe, G. (1830). *A System of Phrenology*. J. Anderson. URL: https://books.google.fr/books?id=ODQSAAAAYAAJ.

Coorebyter, Vincent de (Apr. 24, 2013). "Le retour de la Vieille France." In: *Le Soir*. URL: http://www.crisp.be/2013/04/retour-vieille-france/ (visited on 10/08/2020).

Corbin, Jane (2001). *Under the skin of the BNP*. Documentary about the British National Party, the url links to a full transcript. URL: http://news.bbc.co.uk/hi/english/static/audio_video/programmes/panorama/transcripts/transcript_25_11_01.txt.

Césaire, Aimé (1950/1955). *Discours sur le Colonialisme*. Paris: Présence Africaine. ISBN: 2708705318.

Davidson, Donald (1968). "On Saying That." In: *Synthese* 19.1/2, pp. 130–146. ISSN: 00397857, 15730964. URL: http://www.jstor.org/stable/20114635.

Dénigot, Quentin and Heather Burnett (June 2020). "Dogwhistles as Identity-based interpretative variation." In: *Proceedings of the Probability and Meaning Conference (PaM 2020)*. Gothenburg: Association for Computational Linguistics, pp. 17–25. URL: https://aclanthology.org/2020.pam-1.3.

— (2021). "Using Word Embeddings to Uncover Discourses." In: *Proceedings of the Society for Computation in Linguistics*. Vol. 4. DOI: https://doi.org/10.7275/t4y8-z343. URL: https://scholarworks.umass.edu/scil/vol4/iss1/28.

Derville, Grégory (1997). "La stigmatisation des «jeunes de banlieue»." In: *Communication & Langages* 113.1, pp. 104–117.

Eckert, Penelope (1988). "Adolescent Social Structure and the Spread of Linguistic Change." In: *Language in Society* 17.2, pp. 183–207. ISSN: 00474045, 14698013. URL: http://www.jstor.org/stable/4167922.

— (2008). "Variation and the indexical field1." In: *Journal of Sociolinguistics* 12.4, pp. 453–476. DOI: https://doi.org/10.1111/j.1467-9841.2008.00374.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9841.2008.00374.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9841.2008.00374.x.

— (2012). "Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation." In: *Annual Review of Anthropology* 41.1, pp. 87–100. DOI: 10.1146/annurev-anthro-092611-145828. eprint: https://doi.org/10.1146/annurev-anthro-092611-145828. URL: https://doi.org/10.1146/annurev-anthro-092611-145828.

Epstein, Rob and Jeffrey Friedman (1995). *The Celluloid Closet*. Channel Four Films.

Erk, Katrin (2012). "Vector Space Models of Word Meaning and Phrase Meaning: A Survey." In: *Language and Linguistics Compass* 6.10, pp. 635–653. DOI: https://doi.org/10.1002/lnco.362. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/lnco.362. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/lnco.362.

Fauconnier, Jean-Philippe (2015). *French Word Embeddings*. URL: http://fauconnier.github.io.

Feagin, Joe R., Hernán Vera, and Pinar Batur (2001). *White Racism: The Basics*. 2nd. New York and London: Routledge. DOI: https://doi.org/10.4324/9781003061625.

Feyerabend, Paul (1975). *Against method: outline of an anarchistic theory of knowledge*. London: Verso.

Filimon, Luiza Maria (2016). "From the Dog Whistle to the Dog Scream: the Republican Party's (Ab)use of Discriminatory Speech in Electoral Campaigns and Party Politics." In: *Romanian Journal of Society and Politics* 11 (2), pp. 25 –48.

Findlay, Jamie Y (2017). "Unnatural acts lead to unconsummated marriages: Discourses of homosexuality within the House of Lords debate on same-sex marriage." In: *Journal of Language and Sexuality* 6.1, pp. 30–60.

Firth, John R. (1957). "A synopsis of linguistic theory, 1930-1955." In: *Studies in Linguistic Analysis*. URL: https://ci.nii.ac.jp/naid/10020680394/en/.

Frank, Michael C. and Noah D. Goodman (2012). "Predicting Pragmatic Reasoning in Language Games." In: *Science* 336.6084, pp. 998–998. ISSN: 0036-8075. DOI: 10.1126/science.1218633. eprint: https://science.sciencemag.org/content/336/6084/998.full.pdf. URL: https://science.sciencemag.org/content/336/6084/998.

Franke, Michael (Jan. 2009). "Signal to Act: Game Theory in Pragmatics." PhD thesis. Universiteit van Amsterdam.

Frege, Gottlob (1892). "Über Sinn Und Bedeutung." In: *Zeitschrift für Philosophie Und Philosophische Kritik* 100.1, pp. 25–50.

Garbagnoli, Sara and Massimo Prearo (2017). *La croisade "anti-genre". Du Vatican aux manifs pour tous*. Textuel.

Gärdenfors, P. (2004). *Conceptual Spaces: The Geometry of Thought*. A Bradford book. MIT Press. ISBN: 9780262572194.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou (2018). "Word embeddings quantify 100 years of gender and ethnic stereotypes." In: *Proceedings of the National Academy of Sciences* 115.16, E3635–E3644.

Gazdar, Gerald (1979). *Pragmatics: Implicature, presupposition and logical form*. New York: Academic Press.

Geluso, Joe and Roz Hirch (2019). "The reference corpus matters: Comparing the effect of different reference corpora on keyword analysis." In: *Register Studies* 1.2, pp. 209–242. ISSN: 2542-9477. DOI: https://doi.org/10.1075/rs.18001.gel. URL: https://www.jbe-platform.com/content/journals/10.1075/rs.18001.gel.

Gershwin, George and Ira Gershwin (1937). *Let's call the whole thing off*.

Gilliéron, Jules and Edmond Edmont (1920). *Atlas linguistique de la France, 1902-1910*. Paris: Champion.

Goffman, Erving (1970). *Strategic interaction*. Vol. 1. University of Pennsylvania Press.

Gonen, Hila and Yoav Goldberg (2019). "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them." In: *arXiv preprint arXiv:1903.03862*.

Goodin, Robert E. and Michael Saward (2005). "Dog Whistles and Democratic Mandates." In: *The Political Quarterly* 76.4, pp. 471–476. DOI: https://doi.org/10.1111/j.1467-923X.2005.00708.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-923X.2005.00708.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-923X.2005.00708.x.

Goodman, Noah D. and Andreas Stuhlmüller (2013). "Knowledge and Implicature: Modeling Language Understanding as Social Cognition." In: *Topics in Cognitive Science* 5.1, pp. 173–184. DOI: https://doi.org/10.1111/tops.12007. eprint: https://onlinelibrary.wiley.

com/doi/pdf/10.1111/tops.12007. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12007.

Gratton, Chantal (2016). "Resisting the Gender Binary: The Use of (ING) in the Construction of Non-binary Transgender Identities." In: *University of Pennsylvania Working Papers in Linguistics* 22.2. URL: https://repository.upenn.edu/pwpl/vol22/iss2/7.

Grice, Herbert Paul (1967). "Logic and Conversation." In: *Studies in the Way of Words*. Ed. by Paul Grice. Harvard University Press, pp. 41–58.

Guercio, Nicolás and Ramiro Caso (Nov. 2021). "An Account of Overt Intentional Dogwhistling (forthcoming)." In: *Synthese*.

Gura, Ein-Ya and Michael B. Maschler (2008). *Insights into Game Theory: an Alternative Mathematical Experience*. United Kingdom: Cambridge University Press.

Gutzmann, Daniel and Hans-Martin Gärtner (2013). *Beyond Expressives: Explorations in Use-Conditional Meaning*. Leiden, The Netherlands: Brill. ISBN: 978-90-04-18398-8. DOI: https://doi.org/10.1163/9789004183988. URL: https://brill.com/view/title/24160.

Guénolé, Thomas (2015). *Les jeunes de banlieue mangent-ils les enfants ?* Lormont: Le Bord de l'eau. ISBN: 9782356874177.

Hamilton, William L, Jure Leskovec, and Dan Jurafsky (2016). "Diachronic word embeddings reveal statistical laws of semantic change." In: *arXiv preprint arXiv:1605.09096*.

Haney-Lopez, Ian (2014). *Dog Whistle Politics: How Coded Racial Appeals Have Wrecked the Middle Class*. OUP USA. ISBN: 9780199964277.

Hargreaves-Heap, Shaun P. and Yanis Varoufakis (2004). *Game Theory – a Critical text*. Routledge.

Harris, Zellig S (1954). "Distributional structure." In: *Word* 10.2-3, pp. 146–162.

Henderson, Robert and Elin McCready (2018). "How Dogwhistles Work." In: *New Frontiers in Artificial Intelligence*. Ed. by Sachiyo Arai, Kazuhiro Kojima, Koji Mineshima, Daisuke Bekki, Ken Satoh, and Yuiko Ohta. Cham: Springer International Publishing, pp. 231–240. ISBN: 978-3-319-93794-6.

— (2019a). "Dogwhistles, Trust and Ideology." In: *Proceedings of the 22nd Amsterdam Colloquium*. Ed. by Julian J. Schloder, Dean McHugh, and Floris Roelofsen. Amsterdam, pp. 152–160.

— (2019b). "Dogwhistles and the At-Issue/Non-At-Issue Distinction." In: *Secondary Content: The Semantics and Pragmatics of Side Issues*. Ed. by Daniel Gutzmann and Katherine Turgay. Leiden, The Netherlands: Brill, pp. 222 –245. ISBN: 9789004393127. DOI: https://doi.org/10.1163/9789004393127_010. URL: https://brill.com/view/book/edcoll/9789004393127/BP000009.xml.

— (to appear). *Signaling without Saying: The semantics and pragmatics of dogwhistles*.

Hjelmslev, Louis (1957). "Pour une sémantique structurale." In: *Essais linguistiques*. Ed. by Cercle linguistique de Copenhague. 1959th ed. Copenhagen.

Holy Bible: The New King James Version (1975). Thomas Nelson Publishers. URL: https://www.biblegateway.com/versions/New-King-James-Version-NKJV-Bible/#booklist.

Howard, Jeremy and Sebastian Ruder (2018). "Universal language model fine-tuning for text classification." In: *arXiv preprint arXiv:1801.06146*.

Huber, Gregory A. and John S. Lapinski (2006). "The "Race Card" Revisited: Assessing Racial Priming in Policy Contests." In: *American Journal of Political Science* 50.2, pp. 421–440. DOI: https://doi.org/10.1111/j.1540-5907.2006.00192.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-5907.2006.00192.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5907.2006.00192.x.

— (2008). "Testing the Implicit-Explicit Model of Racialized Political Communication." In: *Perspectives on Politics* 6.1, 125–134. DOI: 10.1017/S1537592708080109.

Hurwitz, Jon and Mark Peffley (Jan. 2005). "Playing the Race Card in the Post–Willie Horton Era: The Impact of Racialized Code Words on Support for Punitive Crime Policy." In: *Public Opinion Quarterly* 69.1, pp. 99–112. ISSN: 0033-362X. DOI: 10.1093/poq/nfi004. eprint: https://academic.oup.com/poq/article-pdf/69/1/99/5393632/nfi004.pdf. URL: https://doi.org/10.1093/poq/nfi004.

J. Hudson & Co. Ltd. (1991). *How To Train Your Dog with the ACME 'Silent' Dog Whistle*. Birmingham.

Jäger, Gerhard (2008). "Applications of Game Theory in Linguistics." In: *Language and Linguistics Compass* 2.3, pp. 406–421. DOI: https://doi.org/10.1111/j.1749-818X.2008.00053.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-818X.2008.00053.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-818X.2008.00053.x.

K. Le Guin, Ursula (1974/2019). *The Dispossessed*. London: Gollancz. ISBN: 9781473228412.

Khoo, Justin (2017). "Code Words in Political Discourse." In: *Philosophical Topics* 45.2, pp. 33–64. DOI: 10.5840/philtopics201745213.

Klemperer, Victor (1947). *LTI - Lingua Tertii Imperii: Notizbuch eines Philologen*. 2nd. Boston, MA, USA: Addison–Wesley.

Korta, Kepa and John Perry (2020). "Pragmatics." In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2020. Metaphysics Research Lab, Stanford University.

Kuo, J.D. (2006). *Tempting Faith: An Inside Story of Political Seduction*. Free Press. ISBN: 9780743287128.

Kusner, Matt, Yu Sun, Nicholas Kolkin, and Kilian Weinberger (2015). "From word embeddings to document distances." In: *International conference on machine learning*, pp. 957–966.

Labov, William (1963). "The social motivation of a sound change." In: *Word* 19.3, pp. 273–309.

— (1966a). *The Social Stratification of English in New York City*. Washington D.C.: Center for Applied Linguistics.

— (1966b). *The linguistic variable as a structural unit*. Washington D.C.: Educational Resources Information Center. URL: https://eric.ed.gov/?id=ED010871.

— (1972). *Sociolinguistic Patterns*. Conduct and Communication vol. 10. University of Pennsylvania Press, Incorporated. ISBN: 9780812210521. URL: https://books.google.fr/books?id=hD0PNMu8CfQC.

Lambert, Wallace E. (1967). "A Social Psychology of Bilingualism." In: *Journal of Social Issues* 23.2, pp. 91–109. DOI: https://doi.org/10.1111/j.1540-4560.1967.tb00578.x. eprint: https://spssi.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-4560.1967.tb00578.x. URL: https://spssi.onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-4560.1967.tb00578.x.

Lamis, Alexander P. (1990). *The Two-party South*. Oxford University Press. ISBN: 9780195065794.

Landauer, Thomas K. and Susan, T. Dumais (1997). "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." In: *Pyschological review* 104 (2). URL: https://doi.org/10.1037/0033-295X.104.2.211.

Lasch, Christopher N. (2016). "Sanctuary Cities and Dog-Whistle Politics." In: *New England Journal on Criminal and Civil Confinement* 16-08. URL: https://ssrn.com/abstract=2748899.

LeMaster, J.R., D.D. Kummings, and W. Whitman (1998). *Walt Whitman: An Encyclopedia*. Garland reference library of the humanities. Garland Pub. ISBN: 9780815318767. URL: https://books.google.fr/books?id=4JuyD59nnYcC.

Leech, Geoffrey N. (1983). *Principles of Pragmatics*. Longman linguistics library. Longman. ISBN: 9780582551107.

Lenci, Alessandro (2018). "Distributional Models of Word Meaning." In: *Annual Review of Linguistics* 4.1, pp. 151–171. DOI: 10.1146/annurev-linguistics-030514-125254. eprint: https://doi.org/10.1146/annurev-linguistics-030514-125254. URL: https://doi.org/10.1146/annurev-linguistics-030514-125254.

Lepore, Ernie and Matthew Stone (2015). *Imagination and Convention: Distinguishing Grammar and Inference in Language*. Oxford Scholarship Online. ISBN: 9780198717188. DOI: DOI:10.1093/acprof:oso/9780198717188.001.0001.

Lewis, David K. (1969). *Convention: A Philosophical Study*. Wiley-Blackwell.

— (1970). "General Semantics." In: *Synthese* 22.1-2, pp. 18–67. DOI: 10.1007/bf00413598.

Lindgren, Elina (2018). "Changing Policy With Words: How Persuasive Words in Election Pledges Influence Voters' Beliefs About Policies." In: *Mass Communication and Society* 21.4, pp. 425–449. DOI: 10.1080/15205436.2017.1406522. eprint: https://doi.org/10.1080/15205436.2017.1406522. URL: https://doi.org/10.1080/15205436.2017.1406522.

Lindgren, Elina and Elin Naurin (2017). "Election Pledge Rhetoric: Selling Policy With Words." In: *International Journal of Communication* 11. URL: https://ijoc.org/index.php/ijoc/article/view/6847.

Livingston, Jennie (1991). *Paris is Burning*. Off-White Productions.

Longhi, Julien (Nov. 2012). "Imaginaires, représentations et stéréotypes dans la sémiotisation du mythe de la banlieue et des jeunes de banlieue." In: *Discours et sémiotisation de l'espace.*

Les représentations de la banlieue et de sa jeunesse. Ed. by Béatrice Turpin. Espaces Discursifs. Harmattan, pp. 123–142. URL: https://halshs.archives-ouvertes.fr/halshs-00940249.

Louise, Déjeans (2017). "L'opposition au Mariage pour tous en France: entre retour du religieux et laïcisation de la religion." In:

Manzini, Thomas, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black (2019). "Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings." In: *arXiv preprint arXiv:1904.04047*.

Mautner, Gerlinde (2016). "Checks and balances: How corpus linguistics can contribute to CDA." In: *Methods of critical discourse studies* 3, pp. 155–180.

McCready, Elin (July 2010). "Varieties of Conventional Implicature." In: *Semantics and Pragmatics* 3.8, pp. 1–57. DOI: 10.3765/sp.3.8.

Mendelberg, Tali (2001). *The Race Card: Campaign Strategy, Implicit Messages, and the Norm of Equality*. Princeton University Press. ISBN: 9781400889181. DOI: doi:10.1515/9781400889181. URL: https://doi.org/10.1515/9781400889181.

Mikolov, Tomas, Kai Chen, Gregory S. Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space." In: *ICLR*.

Miéville, China (2011). *Embassytown*. London: Macmillan.

Montague, Richard (1968). "Pragmatics." In: *Contemporary Philosophy: A Survey, Volume 1*. Ed. by R. Klibansky. La Nuova Italia Editrice, pp. 102–22.

Morris, Charles W. (1938). *Foundations of the Theory of Signs*. University of Chicago Press Cambridge University Press.

Nguyen, Dong (Mar. 2017). "Text as Social and Cultural Data." Massive digital datasets, such as social media data, are a promising source to study social and cultural phenomena. They provide the opportunity to study language use and behavior in a variety of social situations on a large scale and often with the availability of detailed contextual information. However, to fully leverage their potential for research in the social sciences and the humanities, new computational approaches are needed. This dissertation explores computational approaches to text analysis for studying cultural and social phenomena and focuses on two emerging areas: computational sociolinguistics and computational folkloristics. Both areas share the recognition that variation in text is often meaningful and may provide insights into social and cultural phenomena. This dissertation develops computational approaches to analyze and model variation in text. PhD thesis. University of Twente.

Noveck, Ira A (2001). "When children are more logical than adults: experimental investigations of scalar implicature." In: *Cognition* 78.2, pp. 165–188. ISSN: 0010-0277. DOI: https://doi.org/10.1016/S0010-0277(00)00114-1. URL: https://www.sciencedirect.com/science/article/pii/S0010027700001141.

Osborne, Martin J. (2004). *An Introduction to Game Theory*. Oxford University Press.

Parikh, Prashant (2001). *The Use of Language*. Stanford University: CSLI Publications.

— (2010). *Language and Equilibrium*. MIT Press.

Perry, John and Simon Blackburn (1986). "Thought Without Representation." In: *Proceedings of the Aristotelian Society, Supplementary Volumes(* 60.1, pp. 137–166. DOI: 10.1093/aristoteliansupp/60.1.137.

Podesva, Robert J., Jermay Reynolds, Patrick Callier, and Jessica Baptiste (2015). "Constraints on the social meaning of released /t/: A production and perception study of U.S. politicians." In: *Language Variation and Change* 27.1, 59–87. DOI: 10.1017/S0954394514000192.

Pojanapunya, Punjaporn and Richard Watson Todd (2018). "Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis." In: *Corpus Linguistics and Linguistic Theory* 14.1, pp. 133–167. DOI: doi:10.1515/cllt-2015-0030. URL: https://doi.org/10.1515/cllt-2015-0030.

Potts, Christopher (2007). "The expressive dimension." In: 33.2, pp. 165–198. DOI: doi:10.1515/TL.2007.011. URL: https://doi.org/10.1515/TL.2007.011.

Pratt, Mary Louise (1986). "Ideology and Speech-Act Theory." In: *Poetics Today* 7.1, pp. 59–72. ISSN: 03335372, 15275507. URL: http://www.jstor.org/stable/1772088.

Provencher, Denis M (2011). "'I dislike politicians and homosexuals': Language and homophobia in contemporary France." In: *Gender & Language* 4.2.

Rabin, Matthew (1990). "Communication between rational agents." In: *Journal of Economic Theory* 51.1, pp. 144–170. URL: https://EconPapers.repec.org/RePEc:eee:jetheo:v:51:y:1990:i:1:p:144-170.

Rayson, Paul and Roger Garside (2000). "Comparing corpora using frequency profiling." In: *The workshop on Comparing Corpora*, pp. 1–6.

Řehůřek, Radim and Petr Sojka (May 2010). "Software Framework for Topic Modelling with Large Corpora." English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, pp. 45–50.

Reiner, Rob (1987). *The Princess Bride*. Twentieth Century Fox.

Retz, J.-F. Paul de Gondi Cardinal de (1825/1717). *Mémoires*. URL: https://fr.wikisource.org/wiki/M\%C3\%A9moires_(Cardinal_de_Retz).

Rickford, John R. (1985). "Standard and Non-standard language attitudes in a Creole continuum." In: *Language of Inequality*. Ed. by Nessa Wolfson and Joan Manes. 2012th ed. De Gruyter Mouton. ISBN: 978-3-11-085732-0. DOI: doi:10.1515/9783110857320. URL: https://doi.org/10.1515/9783110857320.

Rodman, Emma (2020). "A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors." In: *Political Analysis* 28.1, pp. 87–111.

Ross, Don (2019). "Game Theory." In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2019. Metaphysics Research Lab, Stanford University.

Rudolph, Maja and David Blei (2018). "Dynamic embeddings for language evolution." In: *Proceedings of the 2018 World Wide Web Conference*, pp. 1003–1011.

Récanati, François (2010). *Truth-conditional pragmatics*. Oxford: Clarendon Press.

SOS-Homophobie (2013). *Charte pour un débat parlementaire respectueux*. URL: https://www.sos-homophobie.org/mariage-pour-tous-et-toutes/charte (visited on 12/10/2021).

Safire, William (2008). *Safire's Political Dictionary*. Oxford University Press. ISBN: 9780195343342.

Sahlgren, Magnus (June 2006). "The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces." PhD thesis. Institutionen för lingvistik, Stockholm University. ISBN: 91-7155-281-2.

Saul, Jennifer (2017). "Racial Figleaves, the Shifting Boundaries of the Permissible, and the Rise of Donald Trump." In: *Philosophical Topics* 45 (2). DOI: https://doi.org/10.5840/philtopics201745215.

— (2018a). "Dogwhistles, Political Manipulation, and Philosophy of Language." In: *New Work on Speech Acts*. Ed. by D. Fogal, D.W. Harris, and M. Moss. Oxford: Oxford University Press. ISBN: 9780191059018.

— (2018b). "Immigration in the Brexit campaign : Protean dogwhistles and political manipulation." In: *Media Ethics, Free Speech, and the Requirements of Democracy*. Ed. by C. Fox and J. Saunders. Routledge. ISBN: 9781138571921. DOI: https://doi.org/10.4324/9780203702444-2.

— (June 2019). "What is Happening to Our Norms Against Racist Speech?" In: *Aristotelian Society Supplementary Volume* 93.1, pp. 1–23. ISSN: 0309-7013. DOI: 10.1093/arisup/akz001. eprint: https://academic.oup.com/aristoteliansupp/article-pdf/93/1/1/28846892/akz001.pdf. URL: https://doi.org/10.1093/arisup/akz001.

Schmid, Helmut (1994). "Probabilistic part-of-speech tagging using decision trees." In: *New methods in language processing*.

Schmidgall, Gary (1997). *Walt Whitman: A Gay Life*. EP Dutton.

Scontras, G., M. H. Tessler, and M. Franke (2018). *Probabilistic language understanding: An introduction to the Rational Speech Act framework*. URL: https://www.problang.org.

Scott, Mike et al. (2001). "Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith Tools suite of computer programs." In: *Small corpus studies and ELT*, pp. 47–67.

Searle, John R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Sennet, Adam (2021). "Ambiguity." In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University.

Shakespeare, William (1623/1910). *Twelfth Night or What You Will*. Original work published in 1623. Public Domain. URL: https://en.wikisource.org/wiki/Shakespeare_-_First_Folio_facsimile_(1910)/Twelfe_Night.

Silverstein, Michael (1976). "Shifters, linguistic categories, and cultural description." In:

— (2003). "Indexical order and the dialectics of sociolinguistic life." In: *Language & Communication* 23.3. Words and Beyond: Linguistic and Semiotic Studies of Sociocultural Order,

pp. 193–229. ISSN: 0271-5309. DOI: https://doi.org/10.1016/S0271-5309(03)00013-2. URL: https://www.sciencedirect.com/science/article/pii/S0271530903000132.

Sperber, Dan and Deirdre Wilson (1995). *Relevance: Communication and Cognition*. 1st. Wiley-Blackwell.

Spärck-Jones, Karen (1972). "A statistical interpretation of term specificity and its application in retrieval." In: *Journal of Documentation* 28 (1). URL: https://doi.org/10.1108/eb026526.

Stalnaker, Robert C. (1970). "Pragmatics." In: *Synthese* 22.1/2, pp. 272–289. ISSN: 00397857, 15730964. URL: http://www.jstor.org/stable/20114754.

Stanley, J. (2018). *How Fascism Works: The Politics of Us and Them*. Random House Publishing Group. ISBN: 9780525511847.

Stanley, Jason (2015). *How Propaganda Works*. Princeton University Press.

Szabo, Zoltan Gendler (2006). "The Distinction Between Semantics and Pragmatics." In: *The Oxford Handbook of Philosophy of Language*. Ed. by Ernest Lepore and Barry C. Smith. Oxford University Press, pp. 361–389.

Tadelis, Steven (2013). *Game Theory: an Introduction*. United States of America: Princeton University Press.

Théry, Irene and Philippe Portier (2015). "Du mariage civil au "mariage pour tous". Sécularisation du droit et mobilisations catholiques." In: *Sociologie (online)* 6.1. URL: https://journals.openedition.org/sociologie/2528.

Truan, Naomi (2016). *Débats parlementaires sur l'Europe à l'Assemblée nationale (2002-2012) [Corpus]*. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr. URL: https://hdl.handle.net/11403/fr-parl/v1.1.

Trudgill, Peter (1986). "Dialects in contact." In:

Valentino, Nicholas A., Vincent L. Hutchings, and Ismail K. White (2002). "Cues That Matter: How Political Ads Prime Racial Attitudes during Campaigns." In: *The American Political Science Review* 96.1, pp. 75–90. ISSN: 00030554, 15375943. URL: http://www.jstor.org/stable/3117811.

Valentino, Nicholas A., Fabian G. Neuner, and L. Matthew Vandenbroek (2018). "The Changing Norms of Racial Political Rhetoric and the End of Racial Priming." In: *The Journal of Politics* 80.3, pp. 757–771. DOI: 10.1086/694845. eprint: https://doi.org/10.1086/694845. URL: https://doi.org/10.1086/694845.

Van der Bom, Isabelle, Laura Coffey-Glover, Lucy Jones, Sara Mills, and Laura L Paterson (2015). "Implicit homophobic argument structure: Equal-marriage discourse in The Moral Maze." In: *Journal of Language and Sexuality* 4.1, pp. 102–137.

VanderStouwe, Chris (2013). "Religious victimization as social empowerment in discrimination narratives from California's Proposition 8 campaign." In: *Journal of Language and Sexuality* 2.2, pp. 235–261.

W. Diane Van Zwol (2014). *Poetry Analysis: "Among the Multitude", by Walt Whitman*. URL: https://love2013ad.blogspot.com/2014/12/poetry-analysis-among-multitude-by-walt.html (visited on 11/29/2021).

Wetts, Rachel and Robb Willer (2019). "Who Is Called by the Dog Whistle? Experimental Evidence That Racial Resentment and Political Ideology Condition Responses to Racially Encoded Messages." In: *Socius* 5, p. 2378023119866268. DOI: 10.1177/2378023119866268. eprint: https://doi.org/10.1177/2378023119866268. URL: https://doi.org/10.1177/2378023119866268.

White, Ismail K. (2007). "When Race Matters and When It Doesn't: Racial Group Differences in Response to Racial Cues." In: *American Political Science Review* 101.2, 339–354. DOI: 10.1017/S0003055407070177.

Whitman, Walt (1855). *Leaves of Grass*. Public Domain. URL: https://en.wikisource.org/wiki/Leaves_of_Grass_(1855).

Wilson, Deirdre and Dan Sperber (2002). "Relevance Theory." In: *Handbook of Pragmatics*. Ed. by G. Ward and L. Horn. Blackwell.

Wittgenstein, Ludwig (1953). *Philosophical Investigations*. Wiley-Blackwell.

Woolard, Kathryn A. (1989). *Double Talk: Bilingualism and the Politics of Ethnicity in Catalonia*. Stanford University Press.

— (1991). "Linkages of language and ethnic identity: Changes in Barcelona, 1980-1987." In: *Focus on Language and Ethnicity: Essays in honor of Joshua A. Fishman*. Vol. 2. John Benjamins Publishing.

— (2009). "Linguistic Consciousness among Adolescents in Catalonia: A Case Study from the Barcelona Urban Area in Longitudinal Perspective." In: *Zeitschrift für Katalanistik* 22.

Yu, Liang-Chih, Jin Wang, K. Robert Lai, and Xuejie Zhang (Sept. 2017). "Refining Word Embeddings for Sentiment Analysis." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 534–539. DOI: 10.18653/v1/D17-1056. URL: https://www.aclweb.org/anthology/D17-1056.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang (2019). "Gender bias in contextualized word embeddings." In: *arXiv preprint arXiv:1904.03310*.

de Saussure, Ferdinand (1971). *Cours de linguistique générale*. 3rd. Paris: Payot.

van Rooij, Robert (2004). "Signalling Games Select Horn-Strategies." In: *Linguistics and Philosophy* 27, pp. 493–527.

von Neumann, John and Oskar Morgenstern (1944). *Theory of Games and Economic Behavior*. United States of America: Princeton University Press.

xp0154 (2010). *"Among the multitude" by Walt Whitman*. URL: https://thedifficulty.wordpress.com/2010/10/05/among-the-multitude-by-walt-whitman/ (visited on 11/29/2021).

Åkerlund, Mathilda (2021). "Dog whistling far-right code words: the case of 'culture enricher' on the Swedish web." In: *Information, Communication & Society* 0.0, pp. 1–18. DOI: 10.1080/

1369118X.2021.1889639. eprint: https://doi.org/10.1080/1369118X.2021.1889639. URL: https://doi.org/10.1080/1369118X.2021.1889639.

Part VI

APPENDIX

A

A (VERY) BRIEF INTRODUCTION TO GAME THEORY

This appendix is a presentation of what GT is and a few of its basic principles. It is important to state that I am not a mathematician and have therefore a very narrow understanding of what game theory is, but mathematical objects originally found in game theory have been used in linguistics (and notably pragmatics) for some time now. In this document, I will briefly present the necessary concepts to understand the formal models in this dissertation and nothing more.

This appendix draws inspiration in particular from Benz, Jäger, and van Rooij (2006), Franke (2009), Gura and Maschler (2008), Jäger (2008), Ross (2019), and Tadelis (2013). Any mistake or approximation found in the following pages are my own.

A.1 A (VERY) BRIEF HISTORY OF GAME THEORY

GT as a mathematical framework was born with von Neumann and Morgenstern (1944), but insights about it are a lot more ancient than this. In Ross (2019), we are provided with written examples of game-theoretic thinking dating back to the Spanish Renaissance and even Ancient Greece. This is not surprising, considerations such as those of GT are very interesting to anyone involved in commercial enterprise, war, and really anything that involves strategic thinking (including, of course, games).

Here, however, we are in fact interested in the branch of mathematics.

What is GT?

> *"Game theory undertakes to build mathematical models and draw conclusions from these models in connection with interactive decision-making: situations in which a group of people not necessarily sharing the same interests are requred to make a decision."*
>
> — *Gura and Maschler, 2008*

> *"Game theory is a branch of applied mathematics that models situations of strategic interaction between several agents."*
>
> — *Jäger, 2008*

> *"Game theory has a prescriptive and a descriptive aspect. It can tell us how we should behave in a game in order to produce optimal results, or it can be seen as a theory that describes how agents actually behave in a game"*
>
> — *Benz, Jäger, and van Rooij, 2006*

There are several things at play here, notably:

- GT is a theory of *decision-making*.
- GT is applicable in situations with *several agents*[1].
- GT has a *prescriptive* and a *descriptive* use.

In the context of this work, we are interested in the *descriptive* use of GT, where it is assumed that this mathematical framework can be an appropriate tool for the representations of interaction between several agents, and that its core concepts can be made close enough to phenomena observable in the real world to be efficiently used for the description of these phenomena. This is not necessarily in keeping with other approaches to game theory, which assume that, given a definition of rationality as the maximization of utility (see below), then GT tells its users the *rational way* to act, in spite of their natural inclinations.

At the core, GT is a theory of decision-making, and according to whether you abide by the prescriptive or descriptive approach, it should respectively tell you how agents *should* make decisions or tell you how agents *actually* make decisions. But those agents and decisions are very abstract objects and a few strong assumptions are made with regards to what they are and how they choose. One very important concept behind GT is that of *utility*, and *utility function*.

A.1.1 *Decision theory and* utility

The concept of *utility* stems from insights available in theories about single person choice.

Say you are a person. Say you want to order ravioli at a Chinese restaurant. The menu contains two kinds of ravioli: pork and mushroom. You will have to choose between those two. Maybe you have a strong preference for mushrooms, maybe you have a strong preference for pork, or maybe you have no preference at all, but the idea here is that you can have a *preference relation* over these decisions.

Typically, this preference relation over decisions (let's symbolize it as $\succeq$) is assumed to be *complete*[2] and *transitive*[3]. In other words, for any available decision compared with any other available decision, it is *always* possible to say that you prefer one over the other or that they are equally desirable (completeness) and if one is preferable to another that is itself preferable to another, then it has to be the case that the first decision is preferable to the third (transitivity).

This part here is truly the first step when going from informal decision theory to abstract mathematical models. You may well find that these axioms do not give an accurate picture of what "preferring one thing to another" means, in fact you can probably imagine situations

1 The terms "agent" and "player" will be used interchangeably.
2 Meaning that for a pair of decisions x, y, you can always say $x \succeq y$ or $y \succeq x$.
3 Meaning that if $x \succeq y$ and $y \succeq z$, then $x \succeq z$.

where either one of these axioms do not hold in real life[4], but these have to be the case if we are to build a *utility function*, and without utility function, there is no game theory.

Utility is what you could gain from making a specific decision.

Fundamentally, utilities should reflect *preferences*. A *utility function* takes each decision and assigns it a value in the form of a real number.

Definition A.1.1. Utility function

A utility function $u : X \to \mathbb{R}$, where X is a set of outcomes, is a representation of an agent's preference relation $\succeq$ if for a pair $x, y \in X$, $u(x) \geqslant u(y)$ iff $x \succeq y$.

The idea behind decision theory is that rational agents should choose whatever brings them the highest utility. Defining that utility can however be very tricky, especially when there are many parameters to take into consideration, but the assumption behind decision theory is that there is always a way to build a utility function that reflects your preferences.

But this case is maybe too trivial, a more interesting way to think about decision-making is to add in some degree of randomness. So let's say that you absolutely love mushrooms, much more so than pork. In fact let's say that you love them three times as much and let's arbitrarily define:

- $u(\text{mushroom}) = 15$
- $u(\text{pork}) = 5$

But your stomach does not like mushrooms. So much so that whenever you eat any, there's a one-in-two chance that you'll get sick. Whereas pork is mostly fine[5]. If you get sick, you get no satisfaction from the meal, it is the least preferred of the outcomes, no matter what you chose to eat. Now the situation is a bit different, there are four possible outcomes. We'll represent them in the form of a weighted graph, where the weight of each edge is the probability that an outcome occurs, as in Figure 30.

Now utility in itself is not enough for an agent to make an informed decision. We also need to take into account the possibility of being sick. Once we know that, we have an idea of the *expected utility* of an action, its utility conditioned on the probability of each possible outcome. Adding probabilities into the mix allows us to take into account the relative uncertainty of a situation and to give each decision an expected utility that can be seen as a *score* of sorts.

Definition A.1.2. Expected utility

Let u be an agent's utility function over *outcomes* in a set of outcomes X ; let P be a probability distribution over X such that $p_k = Pr\{x = x_k\}$. The *expected utility* of a decision x given its probability is expressed as:

$EU(x|P) = \sum_{k=1}^{n} p_k \times u(x_k)$

4 Maybe you prefer pork ravioli taste-wise, but mushroom ravioli texture-wise, for example and therefore have to consider two distinct preference relations that contradict themselves, or maybe you think that pork and mushrooms are absolutely impossible to compare.

5 One chance out of ten is maybe not so fine, but it makes things simpler here.

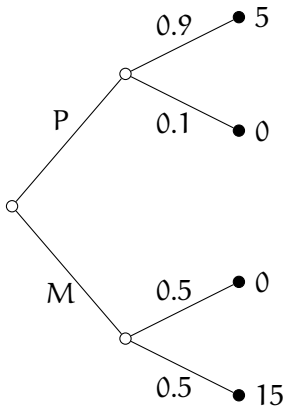


Figure 30: Reprensentation of the outcomes for the ravioli choice.

In the present case, given the situation presented in Figure 30, we therefore have:

$$EU(m) = 0.5 \times 15 + 0.5 \times 0 = 7.5$$
$$EU(p) = 0.9 \times 5 + 0.1 \times 0 = 4.5$$

So in spite of the much higher possibility of sickness, you love mushrooms that much that the rational choice is still to choose the mushroom ravioli[6].

The initial goals of decision theory are as presented here: how do I, a rational agent, make the most rational choice in a given situation. The difference here between a prescriptive and a descriptive reading of decision theory would be as follows:

- The prescriptive approach takes utilities as given, it assumes that we can easily compare all outcomes and their probability, and therefore a set of most rational choices will emerge, and those choices will be those with highest expected utility. It shows you the choice you *ought* to make. This may be very interesting in simple and money-based situation, where the preference function is indeed easier to turn into numbers. It does not exclude the idea of having several agent *types*, but it defines them a priori.

- In the descriptive approach, still assuming that an agent will make the decision that carries with it the highest expected utility, the differences between agents has to be taken into account through the general aspect of the utility function and of the probability distribution over outcomes. It admits as many *rational choices* as there are agent types (which translates in abstraction to utility function types), importantly, a descriptive approach to decision theory forces one to describe agent types a posteriori.

But decision theory is not game theory. It is relevant in simpler cases where there is just one agent considering their options. But there are many cases where the choices made by other agents will matter. When this is the case, we enter the realm of game theory.

6 According to a view of rationality that abides by the rule "Maximize expected utility", and agreeing with the utilities presented here.

A.2 GT IN A FEW CONCEPTS

When many agents are involved and the actions of one agent might influence the outcomes of the others, we speak of game theory, and those decision-making situations are subsequently called "games".

In spite of its name and even though many of its concepts can very much apply to real life games, GT does not only describe what we would commonly call "games". As is often the case in math, games in GT are stripped of most things that make them fun a reduced to a formal object with specific sets of properties.

A.2.1 *Games*

We will not get into the details of GT in this work and will be content with vague, verbal defitions for many terms, because in fact, readers will mostly not need an in-depth knowledge of these concepts to understand what the dissertation is about – the specific games we will be constructing will themselves be formally defined, but a broad grasp of the concepts presented here is most likely sufficient to understand these. Also, I am not a mathematician. If you want to know more, however, Tadelis (2013) is a good reference textbook[7].

With that in mind, a GAME in GT is an abstract representation of a decision-making situation such as the ravioli choice situation described earlier, but involving several agents, where the decisions of each agent are influenced by the decision of the others. Importantly, *games* themselves do not model the reasoning of the agents but only the context in which that reasoning happens. The part of the mathematical models that describes the reasoning[8] is called a SOLUTION CONCEPT, and we will come back to those soon.

Games are the centerpiece of the theory, and game theory is first and foremost an abstract way of representing the situations mentioned above. Because not all situations are equal, not all games will be equal, and distinctions are made between them from a formal point of view. One such distinction that many people usually know about is that between *zero sum games* and *non-zero sum games*. In a few words, zero sum games are games where the sum of the gains and losses at the end of the play are equal to zero. In other words, a player can only win if one or more other players lose. Poker would be such a game, since the sum of the winnings over the entire game are always equal to the sum of the losses over that same game. Non-zero sum games are games that are not like this. For example, all games that rely on cooperation between players to reach the maximum amount of points would be non-zero sum games.

Other important distinctions are those between STATIC and DYNAMIC games and between games of COMPLETE and INCOMPLETE information.

7 Mostly aimed at economists.
8 And where the description/prescription dichotomy is relevant.

PLAYER CHOICES	stay silent	betray
stay silent	-1, -1	-3, 0
betray	0, -3	-2, -2

Table 56: Possible outcomes of the prisonner's dilemma. In each cell, we have a pair of numbers representing each player's utility in this configuration of choices. In this case, the utility is a representation of the number of years spent in prison (seen as a loss). Player 1 is typically the first number and the Row player, Player 2 is the other one.

A.2.1.1 *Static and dynamic games*

A game is said to be STATIC if all the players make their decision simultaneously or if we're in a situation that's tantamount to them making their decisions simultaneously. The most famous example of a static game that's present in most GT resources is that of the PRISONNER'S DILEMMA. It goes as follows:

Two criminals have been caught by the police and placed in separate interrogation rooms. They both risk to go to prison for 1 year for what they were doing when they were caught. They also know that they have information on their partner in crime that could interest the police and reduce their own prison sentence by a year while increasing their partner's by 2 years. They both have the option of *staying silent* or *betraying* their partner. They both know that their partner has the same options as them.

The possible outcomes of such a game can be represented by a matrix such as the one shown in Table 56. The game as it is presented is static because even if players do not literally play simultaneously, they might as well, since the decision is final and there is no way to know the issue until both players have played. From the point of view of players in that game, the fact that they play simultaneously or not is of little relevance, the relevant bit is the time when they *know* about other players' actions, which is the main component of static games.

We can try to sketch out a more formal defintion of static games as follows:

Definition A.2.1. Static game

A static game is a tuple $\langle P, A_{i\in P}, U_{i\in P}\rangle$ where:

- P is a set of players.
- A is a set of actions (one per player).
- U is a set of utility functions (one per player) taking into account outcomes (combinations of actions). So, $u : A_i \times A_j \to \mathbb{R}$.

Static games are usually represented like in Table 56, using a matrix of outcomes[9]. That is one difference that they have with DYNAMIC games, where the matrix representation becomes cumbersome. Dynamic games are games where the players do not play simultaneously. A SIGNALING GAME is such a game, and it has a particular importance in linguistics and pragmatics (as seen in section 2.3).

9 This representation is sometimes called *normal form*.

In a signaling game, the setup is as follows:

A. Player 1 has a *type* that they wish to communicate to Player 2.

B. Player 2 has prior beliefs about the type of Player 1.

C. Player 1 has a set of actions at their disposal that they can do. Each action has a cost, depending on the type of Player 1.

D. Player 2 will update their beliefs about Player 1's type after seeing which action they chose to do.

There are many ways to envision these games, but this is the gist of it: one player (Player 1) wants to communicate an information (their type) to another player (Player 2) using a specific *signal* (one of the actions available to Player 1) that will hint at the information being communicated. An example that is often found to illustrate this focuses on higher education, but we can also think of examples in other domains. Bird mating rituals are a possible illustration of this.

A female bird wants to ensure that the mate that it chooses is sufficiently resourceful and healthy to favor the birth of resourceful and healthy offspring. Birds' communication systems cannot[10] encode that kind of information. Male birds seeking the attention of their female counterparts will therefore have to find a way to signal that they are indeed resourceful and healthy partners. What better way to do it than displays of colors, songs and dances that are very costly in terms of energy and could only be properly performed by resourceful and healthy individuals?

In this situation, a male bird facing a female bird would have the option to either perform or not perform a mating ritual. If the bird is resourceful and healthy, the ritual would be costly, but not as much as if the bird is not resourceful or healthy. The utility of male birds would be a function of their status after making this decision (accepted or rejected as a mate) and the cost of the performance. In all cases, the best outcome for them would be to not perform anything and still find a mate.

For females, the goal is to reproduce exclusiveley with males deemed resourceful and healthy enough. Their utility would be a function of the type of male they are facing and the decision they make of accepting or rejecting said male as a mate. The best outcomes would be all the outcomes where males that are healthy and resourceful are accepted as mates, worst outcomes would be those where unhealthy and non-resourceful mates would be accepted as mates. The decision of accepting or rejecting a male is done after the male has chosen to perform, or not perform, its mating ritual. After observing the ritual, the female should update its beliefs about the male's type accordingly.

This is one possible way of applying signaling games. In all cases, these games are *dynamic* because Player 2's move (belief updating) can only occur after Player 1's move. The fact that

10 As far as we know, of course. There would also be little incentive on the female's side to believe the male if it were to communicate that information.

the decision-making happens in steps makes it so that it is easier to represent those games using a *tree*[11], as in Figure 31.

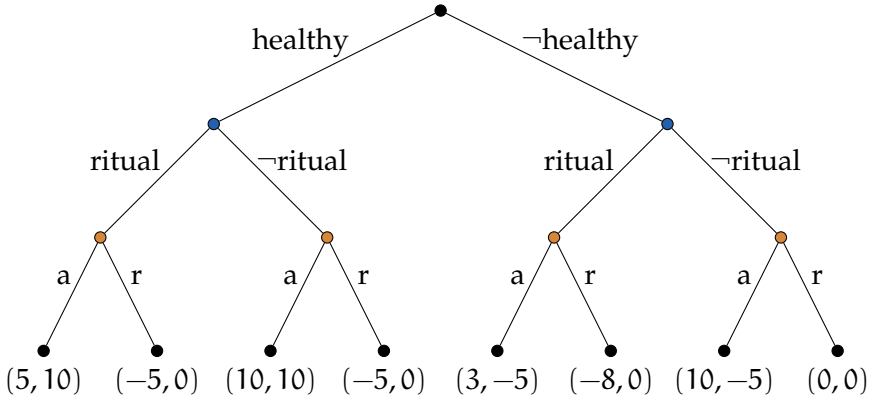


Figure 31: Extended form representation of a signaling game representing mating rituals between birds. The first node represents what is sometimes called *nature*, a player of sorts which decides which type of player Player 1 is. Player 1 moves happen at blue nodes, Player 2's at yellow nodes. "a" stands for "accept" and "r" for "reject". Payoff values for each player are completely arbitrary but should reflect the preferences of each bird in the situation as we have described it.

Because of their significance in pragmatics, a formal definition of signaling games is given in section 2.3.2.

Importantly, both dynamic and static games can be represented in either matrix form or tree form, each form captures different aspects of the games. Matrix forms, while more legible, tend to become cluttered and harder to read when there are several turns to the game. Tree forms more naturally underline the turn-taking that's happening in dynamic games.

A.2.1.2 *Information*

The other dimension according to which games are usually classified is that of *completeness of information*. A game is said to have COMPLETE information when all players know:

A. The actions available to other players as well as their own.

B. The utilities for each outcome for each player.

Coincidentally[12], we've already presented both cases with the prisonner's dilemma and the signaling games. According to this definition, the prisonner's dilemma as it is usually presented is a game of complete information, where all players know the options of other players. Signaling games, on the other hand, are typically games of incomplete information, because the utility for each outcome will depend on the type of Player 1, which is unknown to Player 2.

Note that this is not to be mistaken for PERFECT INFORMATION which refers to games where players have knowledge of their opponents actual moves, and not just their options in terms of

11 This representation is sometimes called *extended form*.
12 Not really coincidentally.

actions. Typically, static games are never games of perfect information, the term rather applies to dynamic games where all moves can be seen by players. Chess would be a game of perfect information in that sense[13].

A.2.1.3 *Dialogue as a game*

Let's take a step back. How is all of this applied to the study of language? Using the game-theoretic terminology that we briefly saw here, we can envision dialogue or conversations as versions of a signaling game. This vision is developed in more detail in section 2.3, but we will present it here briefly as well:

- A Speaker wishes to communicate an information to a Listener.
- The Speaker has a set of messages at their disposal to do so.
- Upon hearing the message chosen by the Speaker, the Listener will have to infer the information that the Speaker intended to send.
- The Listener does not know the type of the information sent in advance, and therefore does not know which interpretation will maximise the Speaker's utility for a given message.

Dialogue is DYNAMIC (Speaker and Listener take turns in a conversation) and it has INCOMPLETE information (there is uncertainty about which interpretation available to the Listener would maximize their own utility as well as the Speaker's).

But as amusing as it can be to try and describe all situations using the terms of GT, the idea behind its elaboration was not just to describe contexts, but also to describe (or prescribe, again) the behaviors of the various agents involved. That is where SOLUTION CONCEPTS come into play.

A.2.2 *Solution concepts*

Any application of game theory should have a good model of the situation in the form of a *game*, but if its goal is to describe *behavior* at least to some extent, then it requires also a good *solution concept*.

First: what is a solution concept? In short, a solution concept is a way to characterize the expected behavior of agents in a game. It is a way for agents to rank the choices that they have according to their preferences and the other players' preferences. Let's see, by way of example, the most famous solution concept of GT.

13 That is, given that you also assume perfect memory for each player.

NASH EQUILIBRIUM The *Nash equilibrium* solution concept is probably the most famous of solution concepts in game theory. Intuitively, a Nash equilibrium is a combination of actions towards which agents will converge after repeated interactions (Franke, 2009)[14]. It is specifically supposed to constitute a steady state in the game from which no one has incentives to deviate. In the case of the prisonner's dilemma of Table 56, there is a Nash equilibrium in which both criminals choose to betray their partner.

The intuition is as follows: even though it might seem that the best possible outcome is the one where no one talks to the police (it has the lowest possible amount of prison years total), individually, each player has an incentive to deviate from that. If I am Player 1 and I assume that Player 2 will not betray me, then I might as well betray them, since I will then be able to leave with no prison time at all. If I assume Player 2 will betray me, then again, I should betray them to reduce jail time. Crucially, Player 2 also knows this. We both have incentives to betray the other, and no incentives to deviate from that position[15].

Equilibria (solution concepts that find steady states in the game) are intuitively interesting when we think about applications of game theory to linguistics, because we want to assume that signal meanings are rather stable over time. It is fairly rare that very unusual usage of a linguistic signal will successfully result in comprehension by listeners.

But a concept like the Nash equilibrium tells us in fact nothing about the reasoning behind the situation where that steady state is reached. If you stick to prescriptive uses of GT, this is sufficient, but in the case of descriptive approaches, it is cognitively lacking, and we would rather look at solution concepts that gradually, in cognitively motivated steps, reach a situation of equilibrium (Nash or otherwise).

PERFECT BAYESIAN EQUILIBRIUM Nash equilibria are applicable to games of complete information, but for games with incomplete information, like signaling games, we have to take into account some degree of belief for the players. An extension of the Nash equilibrium concept to games of incomplete information is the notion of Bayesian Nash equilibrium. The concept is the same except this time each player will choose a strategy that maximizes their expected utility given their prior beliefs about the situation.

So for example, in a signaling game, a Bayesian Nash equilibrium would depend on what the state of the world is thought likely to be by the receiver. If our female bird in the example assumes that the potential mate it's facing is in fact very healthy, then this should come into play in its decision, with or without the execution of a mating ritual.

14 Franke (2009) is here referring to analyses presented in Hargreaves-Heap and Varoufakis (2004) and Osborne (2004).

15 Situations like the prisonner's dilemma, where the outcome that is best for the group is inconsistent with the outcome that is best for each individual player are sometimes called *social dilemmas*, they include situations like the *tragedy of the commons*, sometimes used to explain behaviors of over-exploitation of natural resources. The "no incentives to deviate fomr that position" part can be discussed. Surely, the reduced amount of total jail time in case no one speaks should be an incentive, but in standard approaches to these questions, each player only takes into account their own expected utility. If one wants to model a player that gives more importance to the total amount of jail time than to the amount of jail time that they will be facing, then this should be reflected in their utility function. This is not the starting point assumed in the Prisoner's Dilemma.

Unfortunately this approach can lead to strange and unexpected equilibria in dynamic games like the signaling game. As a response to this, the Perfect Bayesian equilibrium adds the constraint that he play should be optimal starting from any information set. What this means is that whatever strategy you choose should be an optimal strategy whatever information you have and your prior beliefs should be updated at everypossible step of the game that constitutes a subgame[16]. Perfect Bayesian equilibria are often used in the description of dynamic games of incomplete information, like the signaling game.

ITERATED BEST RESPONSE AND RATIONAL SPEECH ACT One framework that gives us more insight on how players can reach equilibria is called the RSA framework (Frank and Goodman, 2012). It is close to another illuminating model, the IBR, a very thorough presentation of which is found in Franke (2009). Because they are central to the work presented here, they are presented directly in section 2.3.

A.3 COOPERATIVE GT AND COOPERATION

Just a few final remarks regarding the notion of cooperation in game theory. A whole branch of the field focuses on cooperative games, notably considering cases of resource sharing, partner attribution and decision-making by vote (Gura and Maschler, 2008).

In the case of applications to linguistics, the word "cooperation" is obviously reminiscent of Grice's CP (see Chapter 2). Indeed in the case of conversation, some degree of cooperation between agents is usually assumed to hold. In the context of signaling games, which we focus on in this work, this has some importance.

The whole concept behind signaling games relies on the idea that the credibility of a signal is underlined by its *cost*. A costless signal (sometimes called *cheap talk*) is not supposed to be a good indicator of the state of the world (or Player 1's type) because anyone could produce it at no cost, and therefore there would be no good reason to believe it. But if the interests of each player are aligned, if it is in fact the case that the best outcome for both players is the accurate communication of information, with no other constraints, then cheap talk can in fact be considered to be credible, since there is no incentive for Player 1 to have Player 2 believe anything other than what the signal conveys (Rabin, 1990, cited in Benz, Jäger, and van Rooij, 2006).

According to Jäger (2008), the CP makes it so that in fact, *all* messages are credible, making it so that costly messages (e. g.longer messages) will generally be avoided in favor of less costly

16 A subgame, is a game within a game. Basically, when you represent a game in extended form, a subgame is any part of the tree that could be construed as a game in itself. Importantly, the initial node of that new game has to be alone in its information set for it to be the initial node of a subgame. This applies to games with imperfect information. In short, if after the move of the first player, the second player does not know which move was just played, then whatever the choice of the first player, the two possible states in which the second player finds themselves are in the same information set. We will not treat games of imperfect information here. The notion of subgame is at the core of *subgame perfect Nash equilibria*, which are one of the main inspirations behind Perfect Bayesian equilibria.

messages. It also means that applications of signaling games to pragmatics inquiry usually ignore message costs.

B

IMPLEMENTATION OF DWG

This appendix presents a short implementation of our model in python which should allow anyone interested to replicate the simulations presented in Chapter 5 and Chapter 6. It can be seen as the documentation to the python program that was used for the computations in those chapters. The code is openly accessible here:

https://github.com/LangdP/DWG

The implementation of the situations found in the dissertation can be found directly on the github page of the project in the form of Jupyter Notebooks, for anyone who would want to check the computations step-by-step. The goal of the present document is to present the content of the program, notably the various classes and methods that are used to implement the model, mostly with the objective of using it. The detail of the code and computations is to be explored directly in the python files at the aforementioned link.

B.1 LEXICA AND PRIORS

Before defining our player classes, we need to define several other objects from the model. The first class we define is the `Lexicon` class, of which `Lex` and `Soc` are subclasses. The corresponding objects in the formal model are Lex and Soc:

```
class Lexicon:
    def __init__(self, utt_dic) -> None:
[...]

class Lex(Lexicon):
    def __init__(self, utt_dic, name="lex") -> None:
        super().__init__(utt_dic)
        self.name = name
[...]

# This is the social meaning lexicon
class Pers(Lexicon):
    def __init__(self, utt_dic, name="soc") -> None:
        super().__init__(utt_dic)
        self.name = name
[...]
```

Objects `Lex` and `Soc` are constructed using an utterance dictionary that has the following form (from `DWG.py`):

```
utterances_nrc = {
    "mR": {"worlds": ["wR"], "personae": ["piRC"]},
    "mNR": {"worlds": ["wNR"], "personae": ["piNRC"]},
    "mDW": {"worlds": ["wNR"], "personae": ["piNRC"]},
}
```

In that dictionary, each key represents an utterance. The corresponding value for each key is another dictionary containing the meaning of the utterance both in terms of "worlds" and "personae". It will be necessary to store all `Soc` and `Lex` in a list when computing the interpretations of listeners.

Another class that is necessary for the creation of players is the `Priors` class. `Priors` are big dictionary-like objects that contain within themselves all the objects of the model that are construed as priors: $\Pr_W, \Pr_\Pi, \Pi$-Lex, Δ-Soc. Each of those priors objects in the model are constructed as some form of dictionary and then are combined together into a `Priors` object (examples from `DWG.py`). Various Soc and Lex objects are referred to by their names, which are defined when constructing `Soc` and `Lex` objects:

```
class Priors:
    def __init__(self, world_priors, pers_priors, delta_soc, pi_lex) -> None:
        self.world_priors = world_priors
        self.pers_priors = pers_priors
        self.delta_soc = delta_soc
        self.pi_lex = pi_lex
[...]

# For listener i
# Define priors over possible worlds here, they have to add up to 1.
world_priors_i = {"wR": 0.5, "wNR": 0.5}

# Define priors over personae here. They have to add up to 1.
pers_priors_i = {"piRC": 0.5, "piNRC": 0.5}

delta_soc_i = {"soc_RC": 1, "soc_NRC": 0}

pi_lex_i = {"piRC": {"lex_RC": 1, "lex_NRC": 0}, "piNRC": {"lex_RC": 0, "lex_NRC": 1}}

# Build priors as an instance of the Priors class.
priors_i = Priors(world_priors_i, pers_priors_i, delta_soc_i, pi_lex_i)
```

B.2 PLAYERS

All the player families are subclasses of the `Player` class, which is constructed using a `Priors` object:

```
class Player:
    def __init__(self, priors: Priors) -> None:
[...]
    def l0_interpretation(self, world: str, utt: str, socs: list, lexs: list)
[...]
    def general_social_interpretation(self, pers: str, utt: str, socs: list)
[...]
    def full_predictions(self, socs: list, lexs: list)
```

The `Player` class has access to several methods that correspond to the computational steps from the model. More precisely, an instance of the `Player` class can be seen as a L_0 listener and has access to all the steps necessary for the computation of $P_{L_0}(w|m)$ and $P(\pi|m)$. These computations are displayed in Figure 2. The predictions of a L_0 listener with any given priors can be displayed using the `full_predictions(socs, lexs)` method and providing the relevant `Soc` and `Lex` objects in the form of lists. The messages are provided as strings. These strings have to be defined in the `Soc` and `Lex` objects that we use for the model to run properly.

B.2.1 *Speakers*

There are 4 classes corresponding to speakers in the implementation. One for each of the speaker families described in Section 5.3.1.2, and an additional one that would stand for a S_2 speaker in a standard RSA model. The `HonestNdivSpeaker` class is a subclass of `Player`. The rest are subclasses of `HonestNdviSpeaker`, which is the equivalent to S_{Reg}.

```
class HonestNdivSpeaker(Player):
    def __init__(
        self, priors: Priors, alpha=1, beta=1, pers_sensitivity=1, world_sensitivity=1)
[...]
        def message_choice(self, utt: str, socs: list, lexs: list)
[...]
        def choice_rule(self, world: str, pers: str, utt: str, socs: list, lexs: list)
[...]

class HonestNdivSpeakerPlus(HonestNdivSpeaker):
    def __init__( self, priors: Priors, alpha=1, beta=1, rank=1, pers_sensitivity=1,
        world_sensitivity=1,)

[...]
```

```
class HonestDivSpeaker(HonestNdivSpeaker):
    def __init__( self, priors_list: list, alpha=1, beta=1, naive_type=0,
        pers_sensitivity=1, world_sensitivity=1)
[...]
    def div_message_choice(self, utt: str, socs: list, lexs: list)
[...]
    def div_choice_rule(self, world: str, pers: str, utt: str, socs: list, lexs: list)

[...]

class DupSpeaker(HonestNdivSpeaker):
    def __init__( self, priors_list: list, worlds_preferences: dict, personae_preferences
        : dict, alpha=1, alpha_bis=1, beta=1, beta_bis=1, naive_type=0, pers_sensitivity
        =1, world_sensitivity=1)
[...]
    def dup_message_choice(self, utt: str, socs: list, lexs: list)
[...]
    def dup_choice_rule(self, worlds: list, perss: list, utt: str, socs: list, lexs: list
        )
```

Each of those classes have a `pers_sensitivity` and `world_sensitivity` parameter. These correspond to the parameters σ, τ mentioned in Section 5.4. The `DupSpeaker` has preferences over worlds and personae in the parameters `worlds_preferences` and `personae_preferences`. We also not that `HonestDivSpeaker` (S_{Div}) and `DupSpeaker` (S_{Dup}) have to take lists of `Priors` as priors, reflecting the fact that they can envision multiple L_0 listeners.

The methods that are of interest to us here are, for `HonestNdivSpeaker`, `message_choice`, corresponding to $P_S(m|w, \pi)$, and `choice_rule`, corresponding to $\mathcal{P}_S(m)$. Corresponding methods exist for `HonestDivSpeaker` and `DupSpeaker`.

B.2.2 *Listeners*

There are 4 classes corresponding to listeners in the implementation. One for each of the listener families described in Section 5.3.1.2, and an additional one that would stand for a L_2 speaker in a standard RSA model. The `Listener` and `ListenerPlus` classes are a subclass of `Player`. The rest are subclasses of `Listener`, which is the equivalent to L_1.

```
class Listener(Player):
    def __init__(self, priors: Priors,  alpha=1,  beta=1,  pers_sensitivity=1,
        world_sensitivity=1)
[...]
    def l1_world_interpretation(self, world: str, utt: str, socs: list, lexs: list)
[...]
    def l1_pers_interpretation(self, pers: str, utt: str, socs: list, lexs: list)
```

```
[...]

class ListenerPlus(Player):
    def __init__(self, priors: Priors,  alpha=1,  beta=1,  pers_sensitivity=1,
        world_sensitivity=1, rank = 1)

[...]

class CageyListener(Listener):
    def __init__(self, priors_list: list, hypothesis_world_prefs: dict,
        hypothesis_pers_prefs: dict, alpha=1, alpha_bis=1, beta=1, beta_bis=1, naive=0)
[...]
    def cagey_world_interpretation(self, worlds: list, utt: str, socs: list, lexs: list)
[...]
    def cagey_pers_interpretation(self, perss: list, utt: str, socs: list, lexs: list)

[...]

class UncovCageyListener(CageyListener):
    def __init__(self, priors_list: list, worlds_pref_priors=dict, pers_pref_priors=dict,
         alpha=1, alpha_bis=1, beta=1, beta_bis=1)
[...]
    def cagey_uncov_world_interpretation(self, worlds: list, utt: str, hyp_pref: str,
        socs: list, lexs: list)
[...]
    def cagey_uncov_pers_interpretation(self, perss: list, utt: str, hyp_pref: str, socs:
         list, lexs: list)
```

We note that `CageyListener` has to be created with hypotheses over world preferences and personae preferences (`hypothesis_world_prefs, hypothesis_pers_prefs`), these preferences come in the form of dictionaries. `UncovCageyListener` has to be created with priors over such preferences (`worlds_pref_priors, pers_pref_priors`), these priors are in the form of a dictionary as well.

The `Listener` class contains the methods `l1_world_interpretation` and `l1_pers_interpretation`, corresponding, respectively, to $L_1(w|m)$ and $L_1(\pi|m)$. The other classes have corresponding interpretation methods.

The `viz.py` file contains functions that allow for the graphical display of the predictions for each player. These functions were used for the generation of the plots in Chapters 5 and 6.

B.3 EXAMPLE

Step-by-step examples are provided in the form of Jupyter Notebooks for several of the examples presented in the dissertation, notably those concerning attempts at replicating RSA and SMG, the "inner cities" example, and the Walt Whitman examples.

The following is the "inner cities" example as it was implemented using the model, which can serve as a reference for how to define each of the necessary terms.

```
# This is the implementation of the DWG model that I describe in my
# dissertation

# Import packages
from players import *
from lexica import *
from helpers import *
from viz import *

# We first have to define the priors for each listener, in the form of two
# dictionaries. The dictionaries are then merged into a Priors object.

# For listener i
# Define priors over possible worlds here, they have to add up to 1.
world_priors_i = {"wR": 0.5, "wNR": 0.5}

# Define priors over personae here. They have to add up to 1.
pers_priors_i = {"piRC": 0.5, "piNRC": 0.5}

delta_soc_i = {"soc_RC": 1, "soc_NRC": 0}

pi_lex_i = {"piRC": {"lex_RC": 1, "lex_NRC": 0}, "piNRC": {"lex_RC": 0, "lex_NRC": 1}}

# Build priors as an instance of the Priors class.
priors_i = Priors(world_priors_i, pers_priors_i, delta_soc_i, pi_lex_i)

# For listener j
# Define priors over possible worlds here, they have to add up to 1.
world_priors_j = {"wR": 0.5, "wNR": 0.5}

# Define priors over personae here. They have to add up to 1.
pers_priors_j = {"piRC": 0.5, "piNRC": 0.5}

delta_soc_j = {"soc_RC": 0, "soc_NRC": 1}

pi_lex_j = {"piRC": {"lex_RC": 0, "lex_NRC": 1}, "piNRC": {"lex_RC": 0, "lex_NRC": 1}}
```

```
# Build priors as an instance of the Priors class.
priors_j = Priors(world_priors_j, pers_priors_j, delta_soc_j, pi_lex_j)

# We then need a set of messages along with their interpretation from a
# lexical standpoint (Lex object) and the social meaning standpoint
# (Soc object).

utterances_nrc = {
    "mR": {"worlds": ["wR"], "personae": ["piRC"]},
    "mNR": {"worlds": ["wNR"], "personae": ["piNRC"]},
    "mDW": {"worlds": ["wNR"], "personae": ["piNRC"]},
}

utterances_rc = {
    "mR": {"worlds": ["wR"], "personae": ["piRC"]},
    "mNR": {"worlds": ["wNR"], "personae": ["piNRC"]},
    "mDW": {"worlds": ["wR", "wNR"], "personae": ["piRC"]},
}

# Constructing lexica and storing in lists

socs = [Pers(utterances_rc, "soc_RC"), Pers(utterances_nrc, "soc_NRC")]
lexs = [Lex(utterances_rc, "lex_RC"), Lex(utterances_nrc, "lex_NRC")]

# Constructing speaker preferences
dw_world_preferences = preferences_generation(
    list(world_priors_i.keys()),
    preferred_states=[["wR", "wNR"]],
    dispreferred_states=[["wNR", "wR"], ["wR", "wR"]],
)
dw_personae_preferences = preferences_generation(
    list(pers_priors_i.keys()),
    preferred_states=[["piRC", "piNRC"]],
    dispreferred_states=[["piNRC", "piRC"], ["piRC", "piRC"]],
)

no_world_preferences = preferences_generation(list(world_priors_i.keys()))
no_personae_preferences = preferences_generation(list(pers_priors_i.keys()))



# Testing

# Literal listeners
L_0_i = Player(priors_i)
```

```
L_0_j = Player(priors_j)

# Vizualize
lis_viz(L_0_i, socs, lexs)
lis_viz(L_0_i, socs, lexs, interpretation="personae_interpretation")

lis_viz(L_0_j, socs, lexs)
lis_viz(L_0_j, socs, lexs, interpretation="personae_interpretation")

# Reg Speaker
S_Reg_i = HonestNdivSpeaker(priors_i)
speak_viz(S_Reg_i, socs, lexs)

# Reg Speaker
S_Reg_j = HonestNdivSpeaker(priors_j)
speak_viz(S_Reg_j, socs, lexs)

# Div Speaker
S_Div = HonestDivSpeaker([priors_i, priors_j])
speak_viz(S_Div, socs, lexs)

# Pragmatic Listeners
Lis_2_i = Listener(priors_i)
lis_viz(Lis_2_i, socs, lexs)
lis_viz(Lis_2_i, socs, lexs, interpretation="personae_interpretation")

Lis_2_j = Listener(priors_j)
lis_viz(Lis_2_j, socs, lexs)
lis_viz(Lis_2_j, socs, lexs, interpretation="personae_interpretation")

#L_2
Lis_2_i = ListenerPlus(priors_i)
lis_viz(Lis_2_i, socs, lexs)
lis_viz(Lis_2_i, socs, lexs, interpretation="personae_interpretation")



Lis_2_j = ListenerPlus(priors_j)
lis_viz(Lis_2_j, socs, lexs)
lis_viz(Lis_2_j, socs, lexs, interpretation="personae_interpretation")

# Duplicitous speaker
S_Dup = DupSpeaker([priors_i, priors_j],
                   no_world_preferences, no_personae_preferences)
speak_viz(S_Dup, socs, lexs)
```

```
# DW Duplicitous speaker
S_Dup = DupSpeaker([priors_i, priors_j],
                   dw_world_preferences, dw_personae_preferences)
speak_viz(S_Dup, socs, lexs)

# Cagey Listener
# Constructing the probabilty distribution on priors for the uncovering cagey
# listener

worlds_prefs_priors = {
    "dw_prefs": {"prefs": dw_world_preferences, "prior": 0.5},
    "npref": {"prefs": no_world_preferences, "prior": 0.5},
}

pers_prefs_priors = {
    "dw_prefs": {"prefs": dw_personae_preferences, "prior": 0.5},
    "npref": {"prefs": no_personae_preferences, "prior": 0.5},
}

L_Cag = CageyListener([priors_i, priors_j],
                      no_world_preferences, no_personae_preferences)

L_Cag_u = UncovCageyListener([priors_i, priors_j],
                           worlds_prefs_priors, pers_prefs_priors)
```

Note that the situations that were considered to be "impossible" by the model in Chapters 5 and 6 will lead to errors when computing the predictions for the cagey listener and the uncovering cagey listener. If one chooses to view the full set of predictions using the `full_predictions` method, errors signaling that such situations exist will arise, but the predictions will be computed anyway, treating those impossible cases as having 0 probability.

C

A (VERY) BRIEF INTRODUCTION TO DISTRIBUTIONAL SEMANTICS

This appendix is a presentation of the very basics of a theory of natural language semantics usually called *Distributional Semantics* and the more recent computational interpretations it has inspired. DS is just one theory of semantics, unlike the kind of semantics that we have tackled in, e. g., Chapter 2, this understanding of semantics is fairly remote from notions like compositionality or interpretations in terms of possible worlds. While the kind of semantics we discussed in Chapter 2 focused mainly on the semantics of full sentences, DS rather focuses on the semantics of words, making it a theory of *lexical semantics*.

We will see what the underlying ideas of DS are and how they have been over the years translated into computational methods that are now standardly used as the basis of many tasks in NLP.

C.1 WHAT IS DISTRIBUTIONAL SEMANTICS?

The works most cited as being the initial inspiration behind DS are Firth (1957) and Harris (1954), and while the field has been explored a lot since then, the basic principles behind the theory are already found in those writings. As a theory, distributional semantics can be seen as relying mostly on two fundamental ideas: one comes from philosophy of language, the other from mathematics.

Regarding the philosophical aspect of the theory, it emerges as truly distinct from formal natural language semantics by asking itself slightly different questions. Semantics as we discussed it in Chapter 2 focuses on the notion of compositionality, and on how the meanings of words come together to build the meanings of sentences, and while great emphasis is put on the definition of the meaning of words in a way that makes them both intuitively understandable (entities are entities), syntactically plausible (adjectives are ultimately sets, i. e., properties) and compositionally justified (adjectives are lambda functions waiting for arguments of entity type), the *meaning* of the word itself in the sense of its reference is defined a priori: the word "dogs" refers to the set of all entities that are dogs. How this particular word came to mean what it means is beyond the scope of traditional formal semantics, so is the potential relationship of this word with other words, like "cats", at least beyond the fact that they are of the same type and will therefore come into the same compositional patterns when deriving the meaning of a sentence. From a formal semantics point of view, "dogs", "cats" and "snakes" are more or less the same thing, meaning that they differ in their reference, but are nonetheless mostly identical.

As language users however, this does not sound like a complete picture. Sure, these words differ in their reference, but they differ differently. "dogs" is more like "cats" than it is like "snakes". And it is more like "snakes" than it is like "computers". Yet there is no overlap in the references of any of these words, from a set theoretic point of view, they are all completely different sets, and there is no good way to quantify that difference[1]. Traditional formal semantics gives us no clue with regards to how the meaning of words emerges, how it evolves, and how it comes to divide the world into the categories that it does. The issue is not raised, and the understanding of semantics it uses does not allow one to attempt to answer this question.

The mechanism proposed by DS is fairly simple, in fact seems trivially true, and is reminiscent of what we said of Wittgenstein's thought in Chapter 2: meaning is use. Wittgenstein focused on a functional understanding of this, reminding of speech act theory; DS focuses on a more surface level interpretation of this statement, seeing the meaning of the word as its use *within* language and not just for the accomplishment of exterior functions in the world. *Usage* here refers, quite literally, to "how the word is used", in what place, in what context, *around which other words*. It is that use that distinguishes word references from one another, makes some words more or less different than others. Any noun that comes after a word like "eat" can therefore be classified in the set of words that come after "eat"; looking at it from a conceptual point of view, those words will likely be considered foodstuff. Let's compare the following sentences:

(35) a. Apples grow on apple trees, they are sweet and a nice snack.

b. Pitayas grow on selinecereus undatus, they are sweet and a nice snack.

c. Pipistrelles grow on pipistrelle trees, they are sweet and a nice snack.

d. $\_\_\__i$ grow on $\_\_\__j$, they are sweet and a nice snack.

Reading those sentences, one might infer that *apples*, *pitayas* and *pipistrelles* are all fruit that grow on trees. In fact one would probably make those same inferences for any of the words that could fill the gaps in (35d). In that particular case, one would be wrong, pitayas are indeed fruit, but they grow on cacti[2], and pipistrelles are a species of bats. The argument still holds, however, and the left and right context provided by the sentences allows one to infer that whatever word appears in gaps i and j are respectively a fruit and whatever it grows on. How the word is used, in the sense of in which sentences, and along with which other words,

1 This can be discussed. One way to see how different the reference of two words are where we would still only use the tools from set theory would be to envision the number of sets that include the references of each word. Then we might see that the reference of both "dogs" and "cats" are subsets of the reference of "mammals", itself a subset of the reference of "animals". If we compare with "snakes", there is only one of those supersets that is shared, and if we compare with "computers", we could probably find shared supersets above "animals", such as "things of the universe", but the smaller sets would not be shared. If we count the number of shared supersets, then this is probably a way to distinguish between degrees of difference while staying in set-theoretic considerations, but one issue with this is that one can always construct such supersets, like we did with "things of the universe". No one stops us from having a word that means "dogs and snakes", vocabulary is known to have oddities like that, see for example the word "pet", which can include just about any animal as long as it is owned for company by a human being. Letting go of the notion of set *per se* is perhaps more fruitful to envision such things.

2 Granted, that's cheating a bit on my part.

is a hint towards its meaning, and if one utters (36) , the effect will be the same, even though "blekturn" isn't a word[3].

(36) Blekturns grow on blekturn trees, they are sweet and a nice snack.

All words that can appear in such contexts will be interpreted as belonging to the same categories of *(edible) fruits* and *fruit-bearing plant*. While the theory does not technically provide us with a cognitive mechanism for how *those specific words* came to refer to *those specific things*, it does provide us with a synchronic reading of *how one classifies new words* and ultimately of *what is "fruit"*, in the sense that the reference of "fruit" will be all those words sharing syntactic contexts that allow us to make a category, which we have chosen to call "fruit".

The mathematical aspect of the theory is linked to the philosophical one: as we have said, the main mathematical theory that is called upon by regular formal semantics *à la* Chapter 2 is *set theory*, leading to a semantics where language discusses *objects*, their *attributes*, or the *relations* that hold between them. There are other areas of mathematics that can however be useful when trying to describe natural language semantics, one such area is *linear algebra*. Why linear algebra? Linear algebra allows us to embed categorical data into a continuous space. This in turn allows the introduction of geometrical notions. One such notion is that of *distance between points*[4], bringing in a flavor of geometry to one's understanding of semantics. And distance is useful to us, because it is exactly the kind of notion that was lacking from traditional formal linguistics approaches when trying to assess the difference between the difference between "dogs" and "cats" and that between "dogs" and "snakes"[5]. Taking a step back from language itself and adopting a conceptual standpoint, linear algebra and vector spaces can satisfyingly convey the idea that objects have properties, and gradual properties with that. For example we could try to take into account two properties, *number of legs* and *number of teeth* and try to see where a number of animals would fall on these scales[6]. See Figure 32 for an illustration. We can see that animals that share traits ($\approx$ same number of legs, for example) are closer in that space, and the more traits are shared, the closer the points are. Ultimately, each animal is reduced to an ordered list of numbers, or *vector*. We could add more relevant traits and make that list longer, which would imply representing each of those points in a space of higher dimension. Beyond 3 dimensions, things become hard to visualize, but the rules of the space remain the same, and there are ways to compute the distance between two points that would allow us to see that concepts that share more characteristics still cluster together.

3 In English. That I know of.

4 Several such notions actually, *cosine similarity*, *Euclidean distance* and *Manhattan distance* are all measures of the distance between two vectors under one understanding of "distance".

5 Yes, that sentence is grammatical.

6 While simple, this example does not technically give us a vector space, but a generalization of a *module*, which is a more general notion of vector spaces. A proper vector space would have to allow for rational values on the axes as well as negative numbers. In this context, however, the approximation is not harmful.

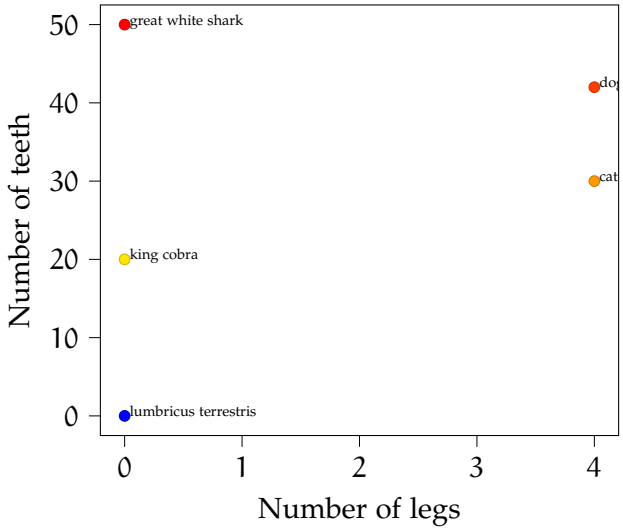


Figure 32: Some animals ranked by number of teeth and number of legs. Yes, the number of teeth on the great white shark is extremely disappointing, but while they do have thousands of teeth over several rows, they apparently only have about 50 that are actively used at any point in their life.

C.2 WORDS AS VECTORS

C.2.1 *How it works*

The core idea behind DS is the exact same, except instead of concepts and properties (which are very much concepts themselves), we talk of *words* and *contexts*.

Note that the definition of *contexts* can be discussed. We could envision it as "the type of text in which a word appears", or "the social situation in which a word is uttered", for example. But what makes DS very interesting, especially from a computational point of view, is the fact that this notion of *context* is understood at its most basic level: which words appear to the left and to the right of the word that we are interested in. This context can be thought of as very wide (all the words appearing alongside the word of interest in a text) or very narrow (just a few words on each side). The size of the context is refined by changing the *window* (the number of words on the left and the right that we take into account), and can be further refined by choosing which words we do take into account (should we take determiners into account? How?), but that is the basic idea. Our words become the points from the visualization in Figure 32, and the possible context words are axes. At the most basic understanding, the position on each axis can be a simple count, but a simple raw counts can cause problems, and more sophisticated measurements are usually preferred, with each dimension of the word vector being rather a score for the corresponding context word.

For example setting a window at 5 words on each side of the word of interest looks like so:

$$\dots w_{-6}, [w_{-5}, w_{-4}, w_{-3}, w_{-2}, w_{-1}, [w_0], w_1, w_2, w_3, w_4, w_5], w_6 \dots$$

The words in blue are the words that will be counted as "occurring alongside" w_0, our word of interest. What do we do with those counts? Unfortunately, using them as is might cause

problems. For example, there are some words that appear *a lot* in any corpus. Because there are many counts for them, we will find them occurring as context words a lot, but they'll appear in the context of most words, and will therefore not be very informative at all. Think about words like "the", or "a", they might co-occur with any of our animal names. One solution would be to simply get rid of such words. This is a solution that can be useful when doing some specific tasks, but it is sometimes very unsatisfactory, because even though words like determiners don't bring in a lot of information from a semantics point of view, they can bring a lot of syntactic information (typically, words that are close to determiners are nouns), which can be useful in some cases.

Another solution, one which is preferred, is to transform those raw counts into *scores* that would reflect the fact that two words are appearing close to each other more often than one would expect if the words were distributed at random. Those scores would be computed using both the counts and the overall frequency of the word in the entire corpus, so that words that appear a lot have to co-occur an abnormally high amount of times with another word in order to make the dimension associated with them go up significantly, while very rare words would be considered highly informative. One operation that allows us to do this is called PPMI (Positive Pointwise Mutual Information); it is defined as follows, for two words w_1 and w_2:

$$\text{PPMI}(w_1, w_2) = \max(\log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}, 0)$$

PPMI will give scores between 0 and $+\infty$. The reason why we use this instead of regular PMI (Pointwise Mutual Information) is that PMI can give negative scores, which may become hard to interpret with bigger corpora; the interpretation of a negative PMI would be that the two words co-occur less than what should happen had the distribution of words been random, which is sometimes an interpretation that is uncalled for.

PPMI is good, but is heavily biased in favor of very rare words. One very easy solution to this is simply to increase the count of every word by n; the higher the n, the less biased towards rarity the count will be, and rarity slowly disappears anyway with all counts becoming increasingly similar in terms of the proportion of total words they represent[7].

Once we have those vectors with PPMI scores instead of raw counts, we can try to compute the difference between any pair of words. We can think of several ways to do so, one measure of distance that is fairly standardly used when comparing word embeddings in NLP tasks is *Cosine similarity*[8], which is computed as follows, for two vectors $\mathbf{A}$ and $\mathbf{B}$ of size n:

7 Think of it this way: if you have a 10 words corpus that goes as follows: "a a a a a a a a a b", then you have 9 "a" and 1 "b", making "a" 90% of the entire corpus. But if you increase the counts for both by, say, 5, then you find yourself with 14 "a" and 6 "b", and suddenly, even though the difference between the two is the same in absolute terms (there are still 8 "a" more), it no longer is in terms of proportions, as instead of "a" representing 90% of the entire corpus, it now represents 70% of the entire corpus.

8 The reason why this is used more often than, say, Euclidean distance, is because it is a measure that is robust to *scaling*. If you have a vector space and you transform it by multiplying each vector in the space by a given scalar (in short, a number by which you will multiply every dimension of the vector), effectively "stretching" or "shrinking" the entire space while keeping relative distances the same, then the Euclidean distance between any two vectors will change, but their cosine similarity will not. In our case, what this means is that it makes it easier to compare words with very differing numbers of occurrences. Words with many occurrences will go further in each

$$S_C(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

We can now compare all the vectors we want. There are other ways to play with those vectors that can have interesting effects however, the most famous of which is undoubtedly the famous example from Mikolov et al. (2013), whereby *king - man + woman = queen*. In the method for the computation of *word embeddings* that is presented in the paper (*word2vec*, which is briefly presented in Section C.3), not only did the clustering of vectors in the space made sense, other operations that are defined on vector spaces, such as vector subtraction and addition, led to interesting results, notably the fact that the vector that was the closest in the space to the vector obtained when taking the vector for *king*, subtracting it the vector for *man* and adding the vector for *woman* was *queen*, leading to the idea that DS done this way could in fact capture deeper semantic traits, like gender. Other similar results are presented there, leading to the idea that the resulting representation has interesting properties for the study of lexical semantics, and that Harris (1954) was indeed on a promising track when proposing the idea.

C.2.2 *In practice*

Let's try with an example.

What does this looks like for a corpus of, say, six sentences? We will take the first two sentences of each of the Wikipedia articles for "Dog", "Cat", and "King cobra"[9]:

[[The [dog] or domestic [dog], (Canis] familiaris or Canis lupus] familiaris) is a domesticated descendant of the wolf which is characterized by an upturning tail. [The [dog] derived from an ancient, extinct] wolf, and the modern grey wolf is the dog's nearest living relative.

[The [cat] (Felis catus) is a domestic] species of small carnivorous mammal. It is the only domesticated species in the family Felidae and is often [referred to as the domestic [cat] to distinguish it from the wild] members of the family.

[The [king cobra] (Ophiophagus hannah) is a venomous snake] species of elapids endemic to jungles in Southern and Southeast Asia. The sole member of the genus Ophiophagus, it is distinguishable from other cobras, most noticeably by its size and neck patterns.

There are here three counts of "dog", and 12 different context words. There are two counts of "cat" and 12 different context words. There is one count of "king cobra"[10] and 7 different

dimension with our method simply due to the fact that there is more of them. To make the comparison between words with fewer occurrences more justified, we use cosine similarity.

9 Respectively available at the following links:
https://en.wikipedia.org/wiki/Dog
https://en.wikipedia.org/wiki/Cat
https://en.wikipedia.org/wiki/King_cobra

10 For the sake of the example, let's treat "king cobra" as a single lexical unit.

context words. The possible context words are all the words in our corpus. Figure 57 shows a subset of the overall *term-term co-occurrence matrix*, a big matrix $\mathbf{M}$ of dimension $m \times m$ with m the number of words in our corpus, and such that $\mathbf{M_{ij}}$ is the number of times word $\mathbf{M_{i*}}$ and word $\mathbf{M_{*j}}$ co-occur given the window size we have chosen. Since most words do *not* co-occur with most other words, the matrix itself is very *sparse*, meaning it contains a lot of zeros; standard practice is to reduce the dimensions of this matrix using techniques such as Singular Value Decomposition (SVD). In the context of this example, we will not do this, if only because our matrix is not very big.

context	dog		cat		king cobra	
a	0	0.00	1	0.33	1	0.45
an	1	0.34	0	0.00	0	0.00
ancient	1	0.41	0	0.00	0	0.00
and	0	0.00	0	0.00	0	0.00
as	0	0.00	1	0.55	0	0.00
...	...	...	...	...	...	...
upturning	0	0.00	0	0.00	0	0.03
venomous	0	0.00	0	0.00	1	0.67
which	0	0.00	0	0.00	0	0.00
wild	0	0.00	0	0.00	0	0.00
wolf	1	0.20	0	0.00	0	0.00

Table 57: Vector representations of the words "dog", "cat" and "king cobra". For each, raw counts are on the left, PPMI scores with one occurrence added for each word are on the right (rounded to two decimals).

Let's try to compare the vectors for "dog", "cat" and "king cobra", using the matrix from Table 57.

$$S_C(\text{dog, cat}) \approx 0.1932$$

$$S_C(\text{dog, king\_cobra}) \approx 0.0162$$

$$S_C(\text{cat, king\_cobra}) \approx 0.0715$$

This is not very impressive, although we can already observe some difference there. To the defense of the method, the corpus we used was extremely small. Here are the results when we take the entire introductions for the 100 first articles in the categories "Dogs", "Cats" and "Snakes". The rest of the parameters are the same: window of 5 and add-one smoothing when computing the PPMI.

$S_C(\text{dog, cat}) \approx 0.5088$

$S_C(\text{dog, king\_cobra}) \approx 0.2052$

$S_C(\text{cat, king\_cobra}) \approx 0.1702$

The tendency is a bit clearer now. Again, when compared with what usually counts as a corpus for such techniques, this is extremely small. We can already see a few interesting things happening, see for example the comparison between "dog" and "wolf" and "cat" and "wolf"; see also the comparison between "cat" and "felis" when compared with "dog" and "felis". Or between "king cobra" and "domestic" when compared with "domestic" and any one of "dog" or "cat". Though we are working on a still fairly small corpus, we can already capture interesting intuitions.

$S_C(\text{dog, wolf}) \approx 0.4227$

$S_C(\text{cat, wolf}) \approx 0.3024$

$S_C(\text{dog, felis}) \approx 0.2318$

$S_C(\text{cat, felis}) \approx 0.3526$

$S_C(\text{domestic, king\_cobra}) \approx 0.1497$

$S_C(\text{domestic, dog}) \approx 0.4384$

$S_C(\text{domestic, cat}) \approx 0.5306$

Interestingly, because of how cosine similarity is defined, the way that syonymy or more generally semantic similarity is defined in DS is a continuous measure that applies to all pairs of words. Meaning that there is no such thing here as "incomparable words", you can always find some similarity score for any two pairs of words; but for most cases, those scores will approach 0, and only the bigger ones might have something to tell[11]. One interesting thing to do then is to look at the set of the n most similar words to any given word. Those are the words whose vectors are, in the vector space we have created, closer to the vector of the word of interest, the words that have "clustered together".

The matrix generated here containing all the vector representations of our words is called a *vector space model*, and the vector representations themselves are sometimes called *word embeddings*.

The method can be optimized if we want to have more specific results, the main parameter that we can play with here is the window size. There has been work, presented in Sahlgren

11 And even then, if the distribution of words is very unequal, things will be even harder to interpret.

(2006), to try and identify concepts from DS with traditional structual linguistics terms, which has led notably to the general understanding that the size of the window can greatly influence the results. For example, a common understanding of word embeddings is that smaller window sizes (say, < 5) will tend to lead to the most similar words to a given word being words typically substitutable on the paradigmatic axis (so typically, words most similar to nouns would be other nouns), whereas larger windows would lead to a different kind of semantic similarity, including notably meronyms of hyponyms of the word of interest. So for example, a smaller window should make it so that "cat" is fairly close to "dog", while a larger window would make it so that "dog" is closer to, say, "snout" or "leash". This is interesting in the sense that it underlines that the notion of semantic similarity that is put forward by DS is in fact compatible with several understandings of "semantic similarity".

All of this will nonetheless be highly dependent on the corpus that the VSM is built on, and results may differ due to many properties of said corpus.

C.3 *word2vec*

There are several computational methods that rely on the ideas of DS and can give us vector representations for the words in a corpus (two very influential tools are TF-IDF, first presented in Spärck-Jones, 1972, and LSA, in Landauer and Dumais, 1997). The one that we have used in the context of this work is called *word2vec*. It was first described in Mikolov et al. (2013) and has since become very standard, but at the time when it came out, neural network based approaches to this task were not as common as they are now. We will present it very briefly and see how it differs from the very basic picture of DS we've given so far.

Neural networks are typically used for their *predicting* abilities, meaning that what word2vec fundamentally does is try to *predict* something. It is through this predicition task, its successes and failures, that the vectors for our words are created, and not merely through a word count like we did earlier. In a way, the neural network approach is closer to the philosophical understanding of DS; the initial intuition for DS comes from the intrinsic ability to guess which words can go in a sentence with a missing word, and which words are likely to appear alongside a given word. The neural network approach to this trie to emulate exactly that.

Specifically, word2vec has two distinct implementations:

1. *Skip-gram* (SG), where the network uses each word as an input and predicts the context around it, based on the size of the window.

2. *Continuous bag-of-words* (CBOW), where the network uses the context in the window as an input and predicts the word at the origin of the window.

Like what we saw in earlier sections, the idea behind word2vec is to use a corpus and represent it as a vector space where each word is a vector of n dimensions. Note that word2vec,

unlike what we did just before, does use some form of dimension reduction, allowing one to choose the size of the vectors, a parameter that can have its importance.

The algorithm starts with *one-hot encodings* of each word in the corpus, meaning that each word has a starting vector of dimension V, with V the number of unique words in the corpus, that is filled entirely with 0 except for one specific dimension, specific to that word, which contains a 1. So for example, we could have something like the following:

$$\begin{array}{cccc} \text{a} & \text{aardvark} & \cdots & \text{zymurgy} \\ \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} & \cdots & \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \end{array}$$

This is more or less what the algorithm takes as a starting point. These one-hot vectors are then multiplied by a *weight* matrix $\mathbf{W_1}$ of size $V \times N$ with N the size of the output vectors that we want. The values in this weight matrix are initially set randomly[12]. According to which approach one chooses, what happens then differs:

SG If one chooses the skip-gram approach, what happens is that is either a single one-hot vector serves as the input to the model, that of the word at the center of the window, leading after multiplication with the weight matrix $\mathbf{W_1}$ to a single vector of dimension N which is then used for the prediction of several words. In order to do this, the vector is multiplied by a second weight matrix $\mathbf{W_2}$ of dimension $N \times V$, containing the representations of the context words. That multiplication is repeated several times, once for each of the context words that have to be predicted. The resulting vectors are of size V. These are in turn transformed into probability distributions via softmax, and these probability distributions are compared with the one-hot representations of the words that those were supposed to predict. The difference between the one-hot encodings and the predicted distributed representations constitutes the loss function the model. In short, for the SG model, the input is a single one-hot vector, the output is a set of probability distributions over one-hot vectors based on some transformation of the initial one-hot vector. The discrepancies between that distribution and the actual words of the context around the word that gave us the input vector constitute the loss; using gradient descent, the algorithm modifies the weights in $\mathbf{W_1}$ to minimize the loss. Figure 33 illustrates this.

12 There is a very good reason for this that is beyond the scope of this introduction, but is essentially that one cannot know in advance whether the loss function of the neural network will have a single minimum point that is easily accessed through gradient descent. Setting the initial weights randomly can allow one to have different results when starting a new model with a corpus already used and is in the end the most economical way to minimize the chance of getting stuck in a local minimum that would be impossible to escape even across models would the algorithm start with a definite set of initial weights.

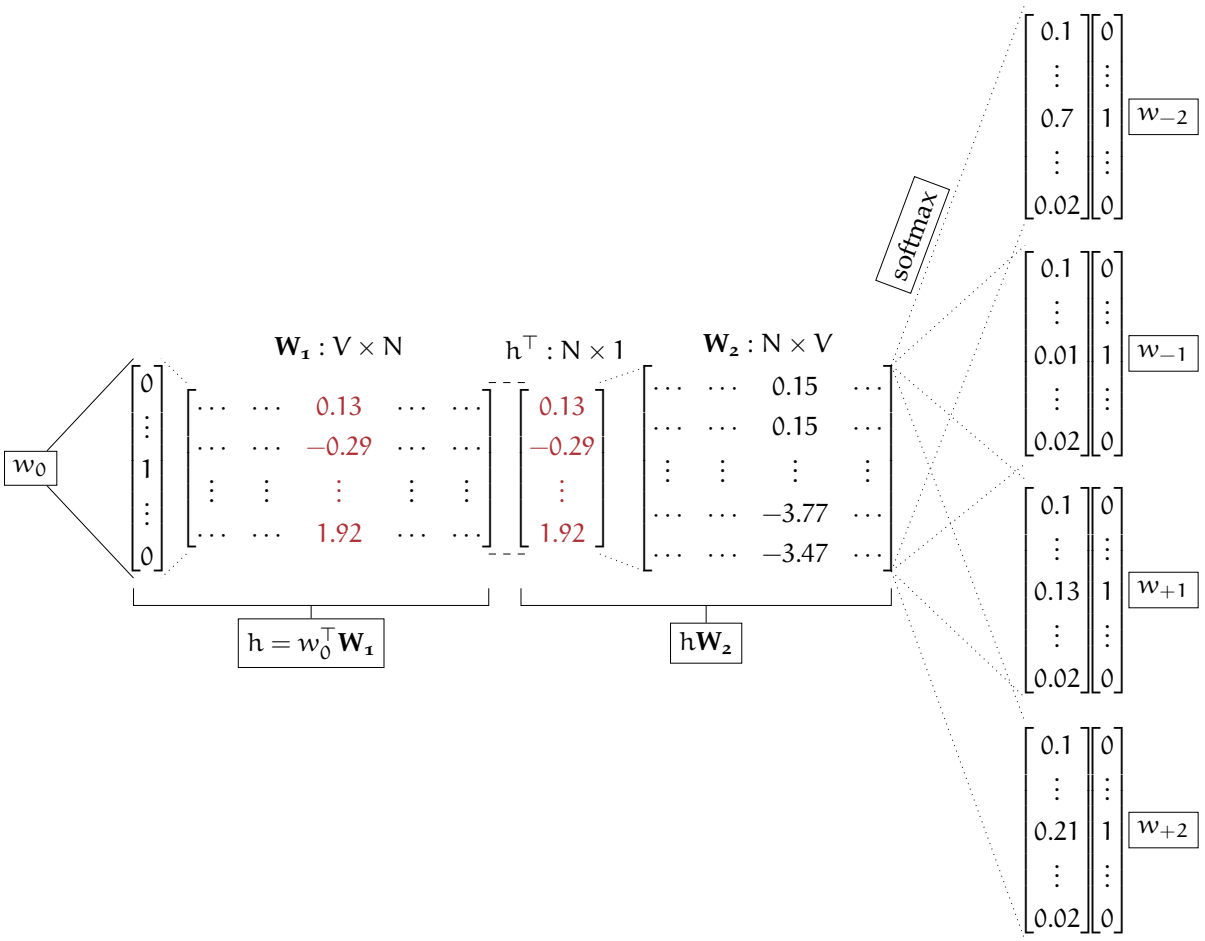


Figure 33: Overview of the skip gram model for word2vec. The input is a single one-hot vector representation of a word, the output is a probability distribution over one-hot vectors. The loss in prediction (see for example w_{-1}, wrongly predicted to be less likely than other words) is used for the updating of the weight matrices. Note how the first multiplication leads to what is effectively the distributed vector representation of our word (highlighted in red).

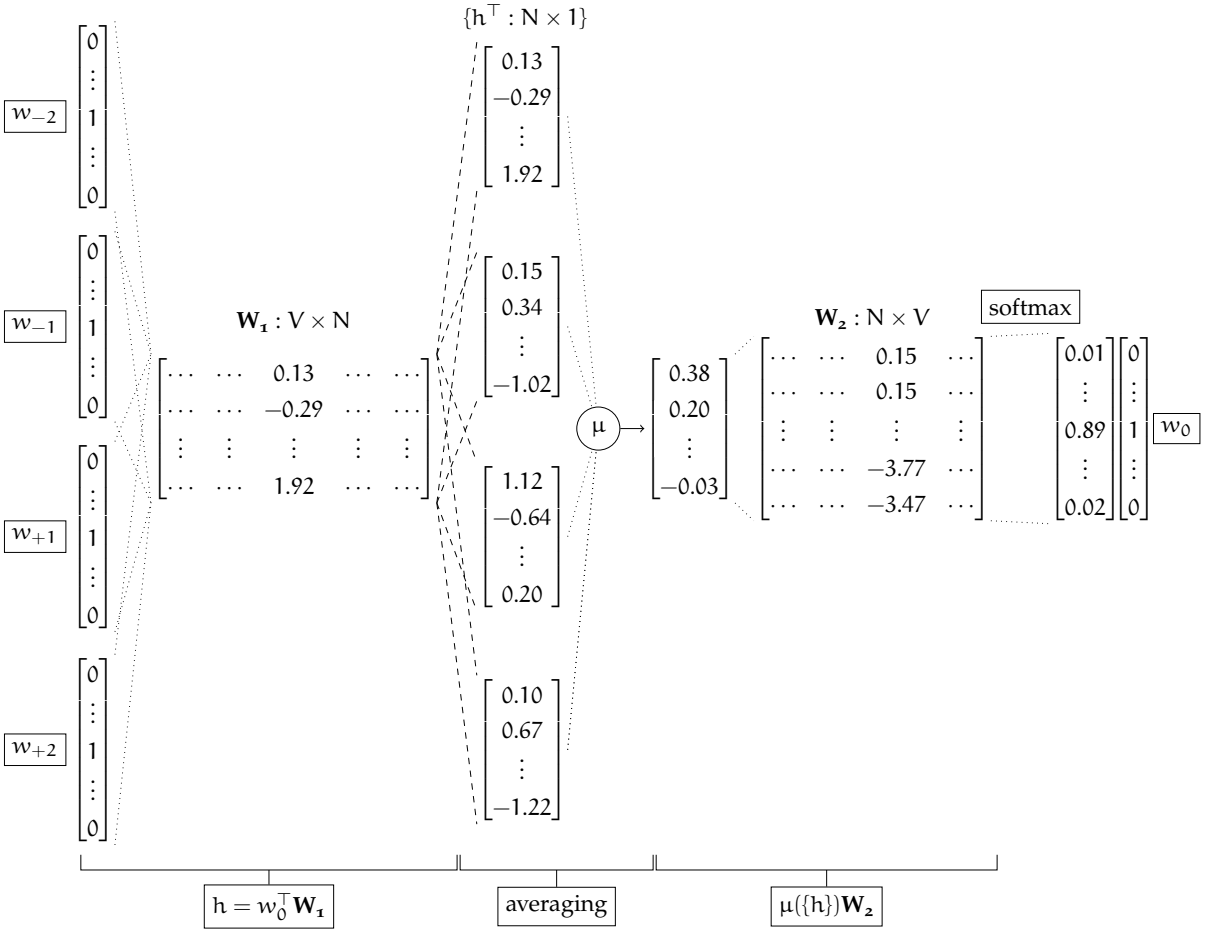


Figure 34: Overview of the continuous bag of words model for word2vec. The input is a set of one-hot vectors, the output is a probability distribution over one-hot vectors. The loss in prediction is used for the updating of the weight matrices. The set of word vectors that serve as input are averaged (dimensions of $\mu : N \times 1$) after being transformed through matrix multiplication with $\mathbf{W_1}$. The resulting vector is multiplied with $\mathbf{W_2}$.

CBOW If one chooses the continuous bag-of-words approach, what happens is that or it can be that several one-hot vectors are the input to the model, are each multiplied by the weight matrix and that all of the resulting vectors of size N are averaged into one single vector. Each of the dimensions of this vector (obtained by element-wise computation) then goes through a second weight matrix $\mathbf{W_2}$ that turns it back into a vector representation of size V, which is turned a probability distribution. This resulting vector is compared with the one-hot representation of the word that is actually present in the text, and the difference between the two representations constitutes the *loss* of the model. That loss is then used to update the weights from $\mathbf{W_1}$. This is repeated with a fixed window size over the entire corpus, with each new input updating the weight matrix, using *gradient descent* with the hope of minimizing the loss. Again, in short, for the CBOW model, the input is a set of one-hot vectors, the output is a probability distribution over one-hot vectors based on some transformation and combination of the initial vectors. The discrepancies between that distribution and the observed word at the center of the window that gave the input vectors constitutes the loss; using gradient descent, the algorithm modifies the weights in $\mathbf{W_1}$ to minimize that loss. Figure 34 illustrates this.

The two models can be seen as symmetrical to some extent: one uses one word to predict many words, the other uses many words to predict one word. And what happens then? What the algorithm does is changing the values of the matrix $\mathbf{W_1}$, which is itself a series of vectors of dimension N, one for each of our words. In other words, the weight matrix $\mathbf{W_1}$ *is* the vector space representation of the words in our corpus. The prediction task in itself is not the interesting part, it is just a means to the actual end of updating the representations of the words in the vector space.

Actual implementations of word2vec use tweaks to make the computations more efficient, things like *negative sampling*, but we will not go into these details here. The implementation that was used in the context of this work is that found in the *gensim* library (Řehůřek and Sojka, 2010).

C.4 CONCLUDING REMARKS

This appendix was just a short introduction to the topic of DS, it is by no means exhaustive, and interested readers are invited to read the many (many) existing introductions and overviews to this and related topics, e. g., Boleda (2020), Clark (2015), Erk (2012), Lenci (2018), and Sahlgren (2006).

By way of conclusion, there are a number of reasons why DS is interesting to us in the context of the study of dogwhistles:

1. Our vision of dogwhistles is centered on a notion of semantic similarity/distance, specifically a differentiation in semantic distance between two words across communities. DS provides us with a clear definition of what *semantic similarity* might mean, and a way to measure it.

2. In general, DS and WE are presented as deriving semantic information from purely syntactic input. This means that the ideas put forward by DS allow one to gather some semantic insights, including isolable traits like *gender* with no information beyond what is directly found in the signal (here, the text). There is no a priori information about words, there is no supplementary semantic information, word categories and groups of similar words simply emerge from the raw data. From a computational perspective, this makes DS and related theories particularly adequate for use with computers and automatic analysis of data.

We should keep in mind however that DS is just one way of envisioning the semantics of words, and that it is neither complete nor entirely satisfactory for some questions. There are good reasons to think that the distributional approach in and of itself is insufficient to characterize sometimes fundamental differences between words. For example, the adjectives "good" and "bad" share a lot of their contexts, but they have meanings generally thought of as opposite. The distributional hypothesis cannot account for this. If we equate the semantics

of a word to its position in a space, or rather to the identity of the words that surround it in a space, and that this space is constructed exclusively using information on the distributions of words, then we do no have a full picture of the semantics of a word.

What we do have is a computationally implementable picture of semantics, and more importantly, one that can rely exclusively on observable data.

Chapters 7 and 8 provide a possible illustration of how these techniques can be thought to be adapted to the task of detecting and studying dogwhistles.

D

SUPPLEMENTARY DATA FOR CHAPTER 8

This appendix presents supplementary data for the attempt at using word embeddings for the exploration of the MPT and PACS corpora presented in Chapter 8. The layout of the plots is the same as Figure 13. The plots already presented in Chapter 8 are not reproduced here.

D.1 MPT

D.1.1 *Test words*

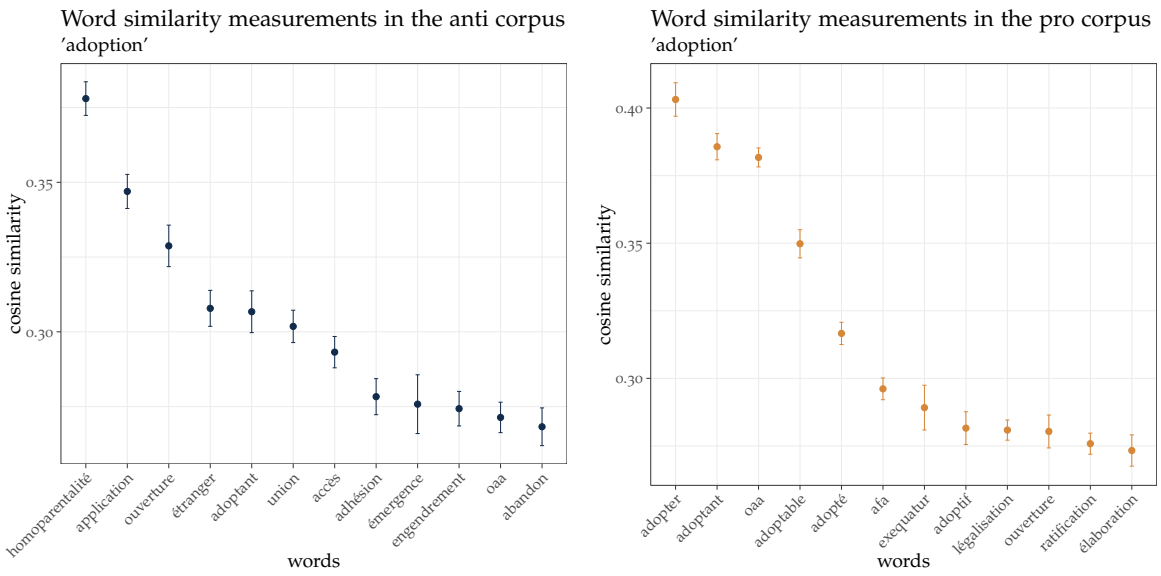


Figure 35: Comparison of closest semantic neighbours for *adoption* across MPT models.

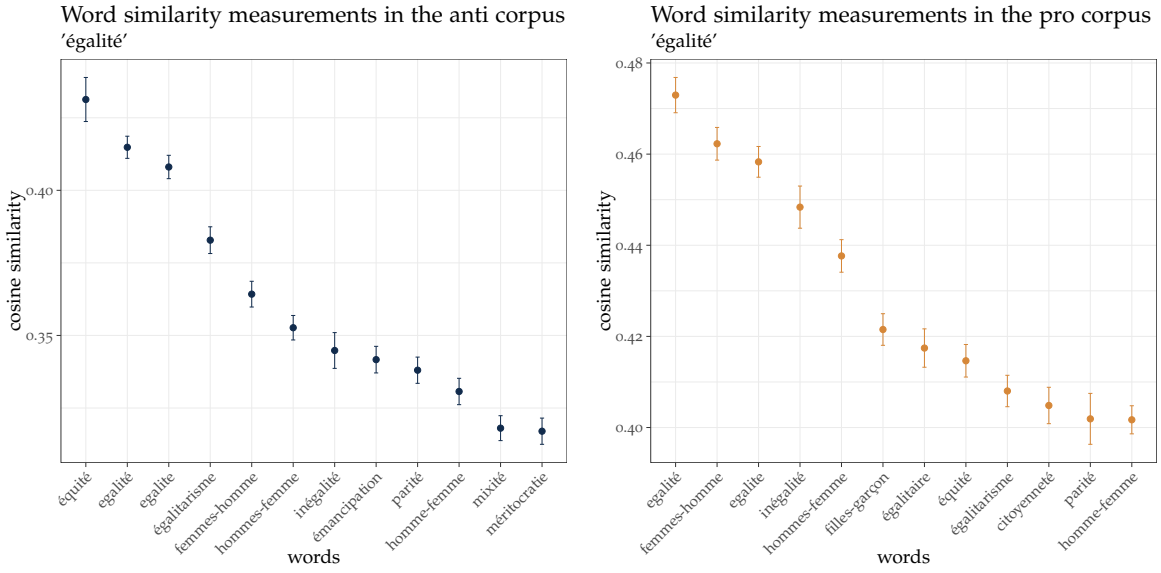


Figure 36: Comparison of closest semantic neighbours for *égalité* across MPT models.

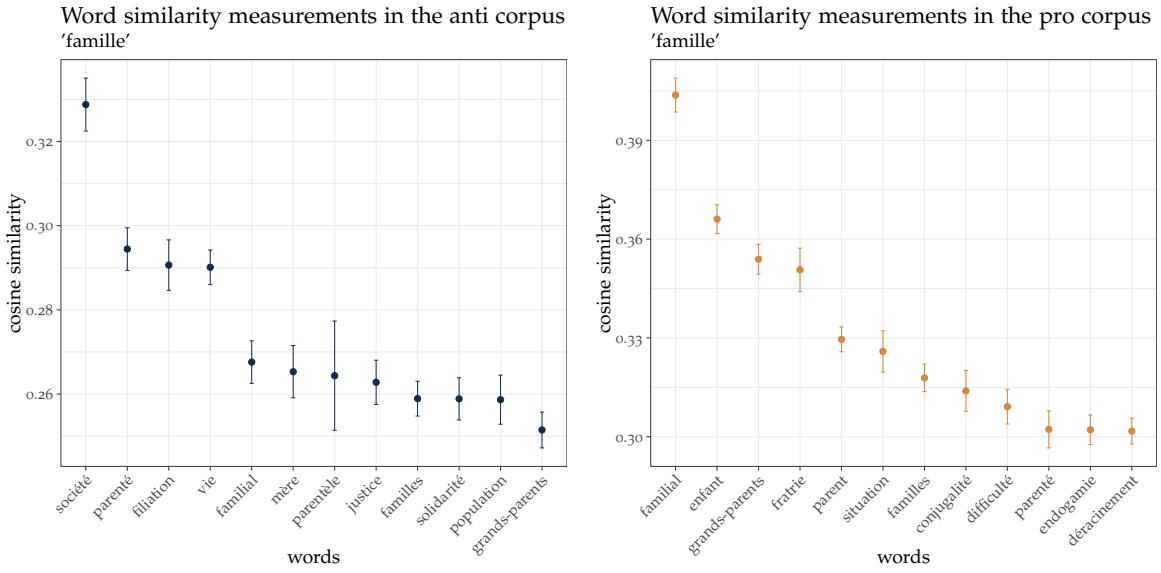


Figure 37: Comparison of closest semantic neighbours for *famille* across MPT models.

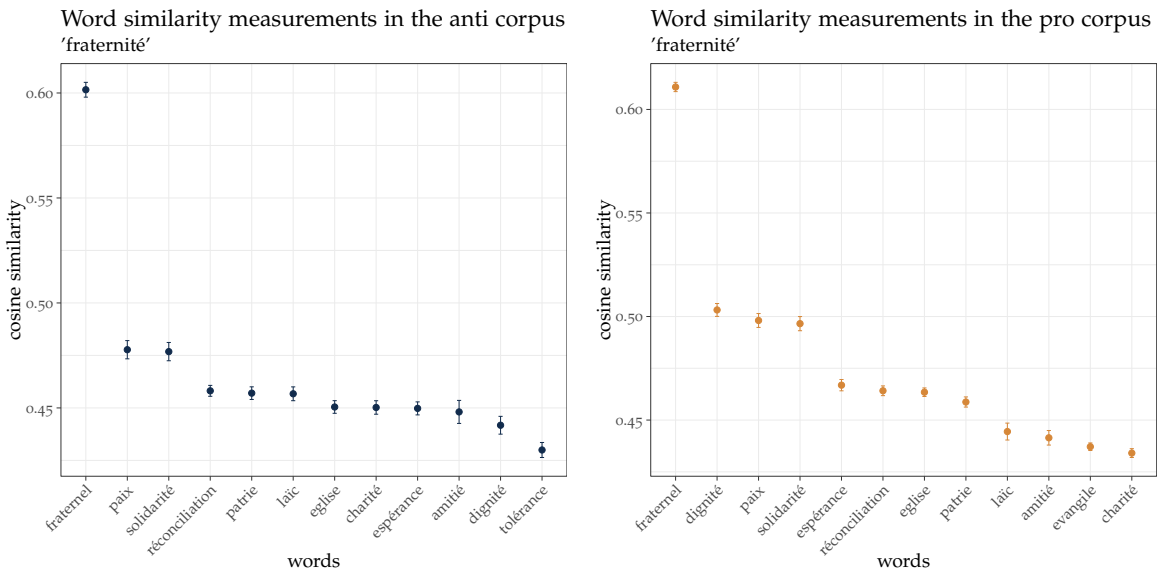


Figure 38: Comparison of closest semantic neighbours for *fraternité* across MPT models.

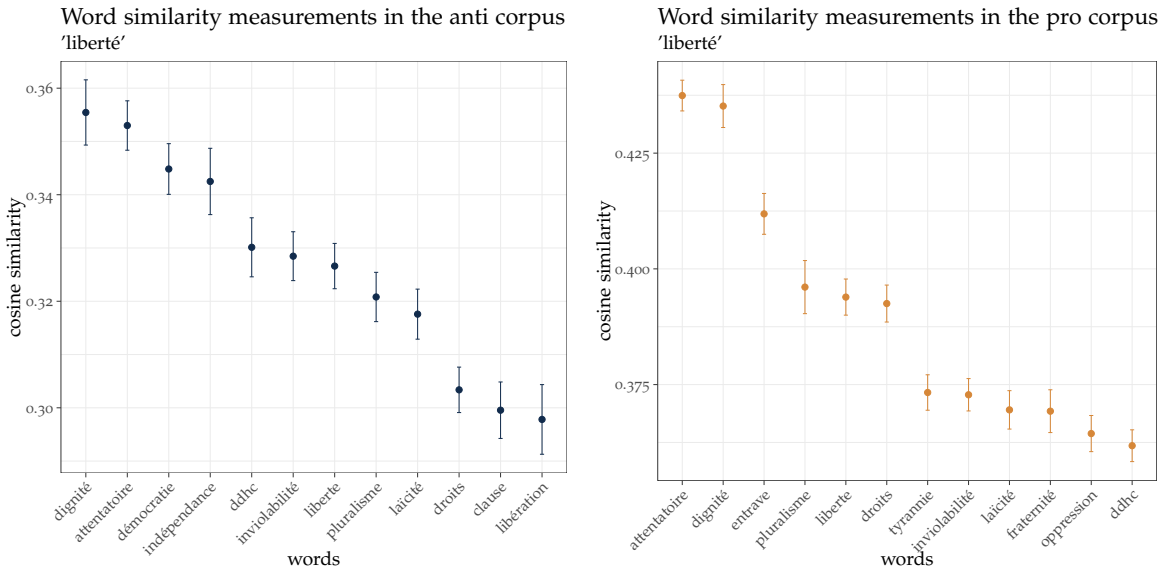


Figure 39: Comparison of closest semantic neighbours for *liberté* across MPT models.

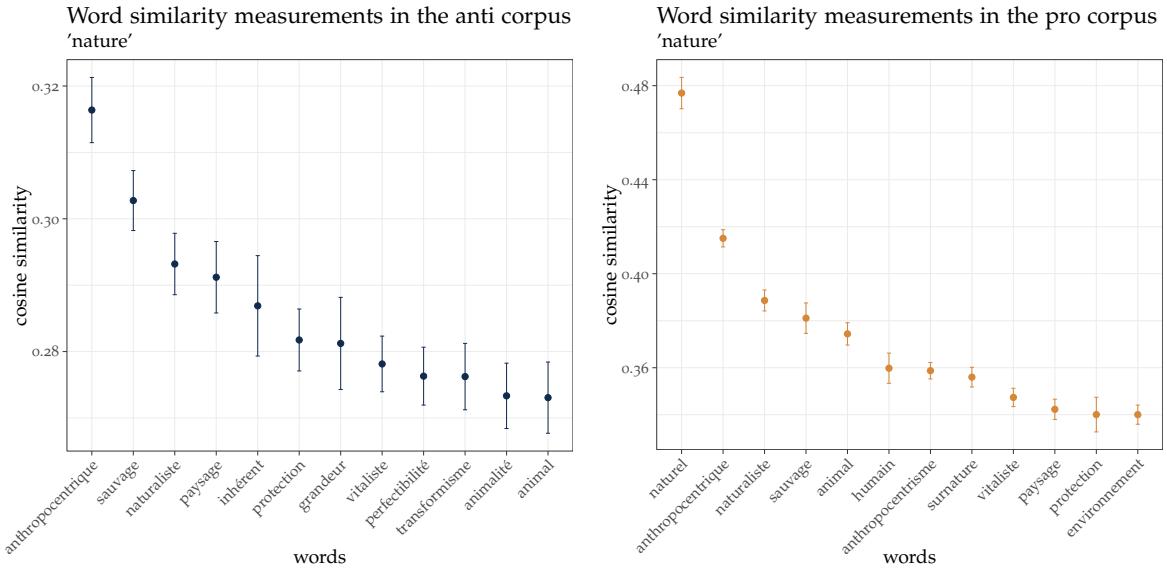


Figure 40: Comparison of closest semantic neighbours for *nature* across MPT models.

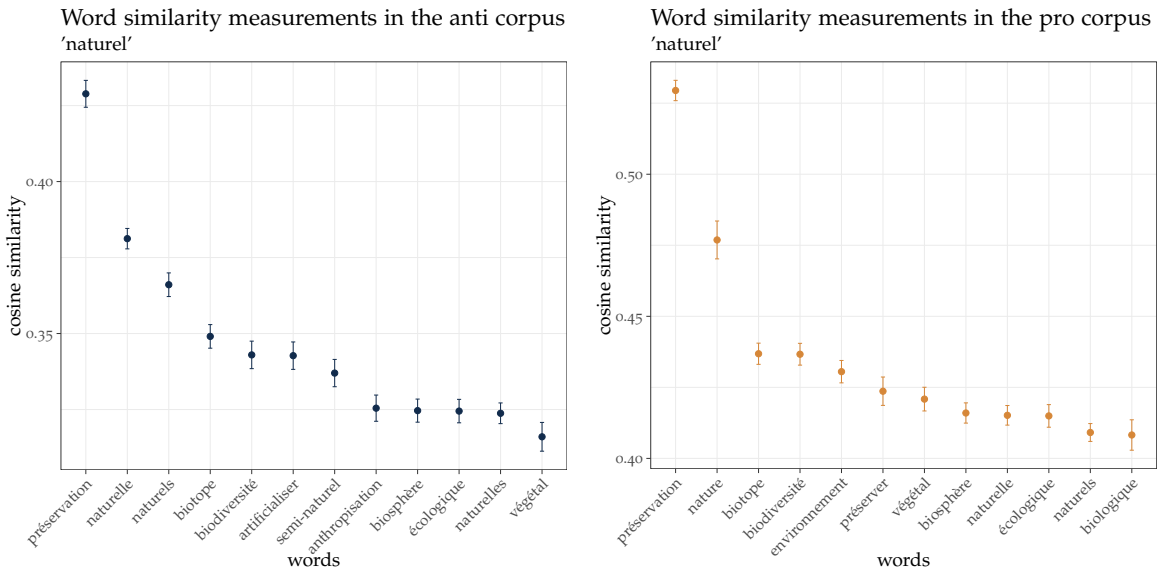


Figure 41: Comparison of closest semantic neighbours for *naturel* across MPT models.

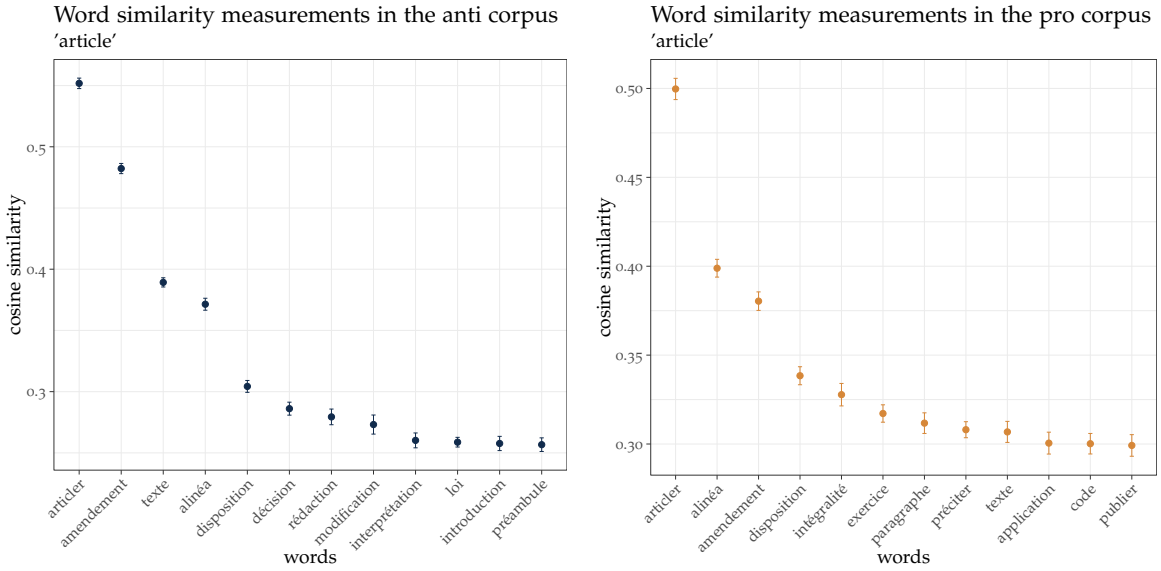


Figure 42: Comparison of closest semantic neighbours for *article* across MPT models.

D.1.2 *Control words*

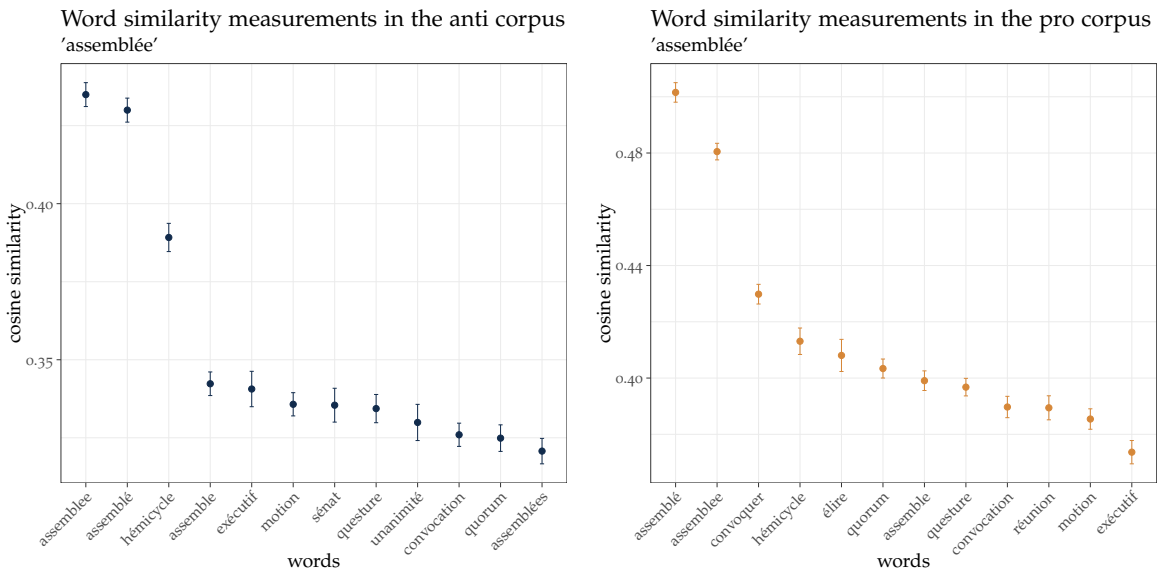


Figure 43: Comparison of closest semantic neighbours for *assemblée* across MPT models.

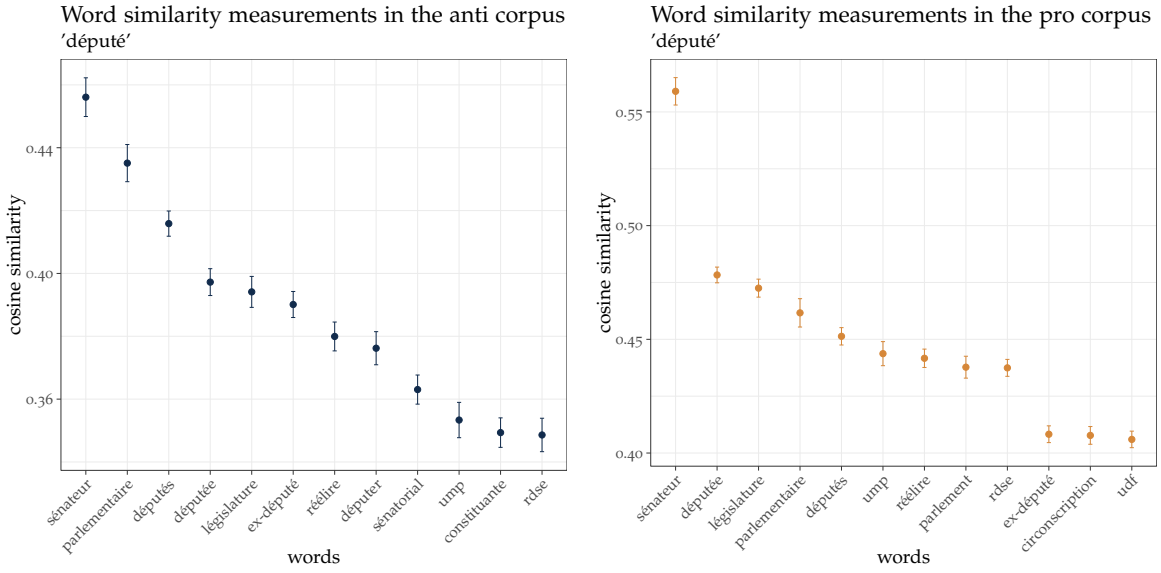


Figure 44: Comparison of closest semantic neighbours for *député* across MPT models.

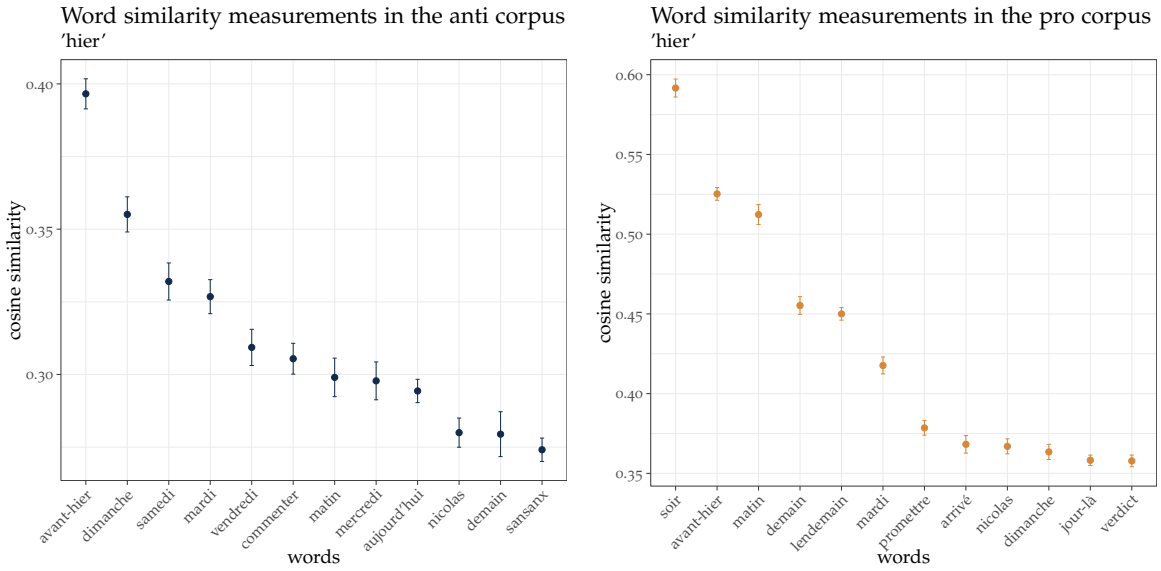


Figure 45: Comparison of closest semantic neighbours for *hier* across MPT models.

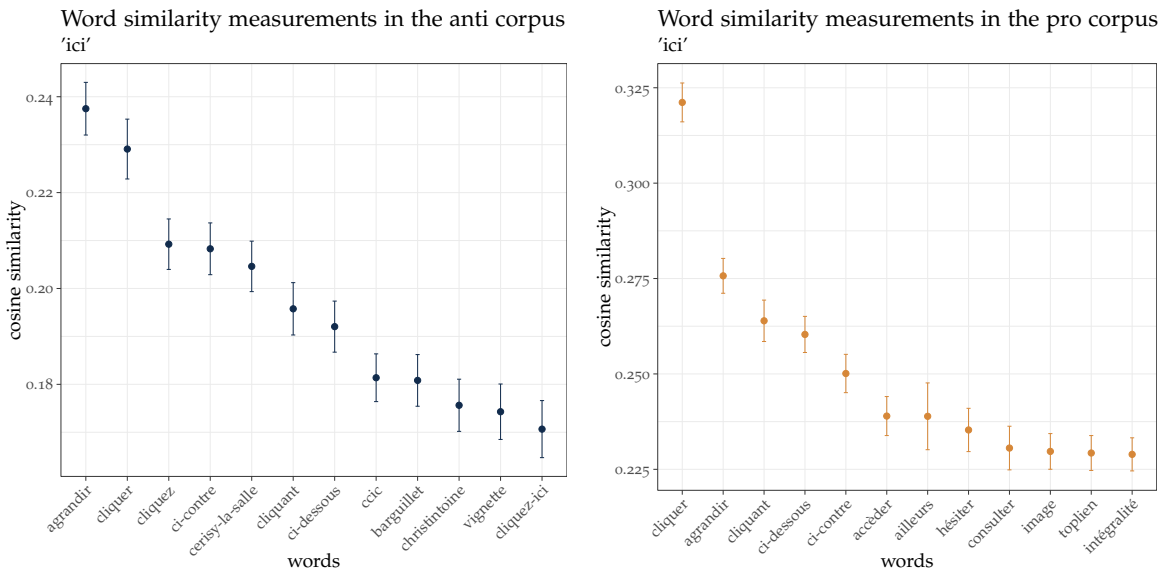


Figure 46: Comparison of closest semantic neighbours for *ici* across MPT models.

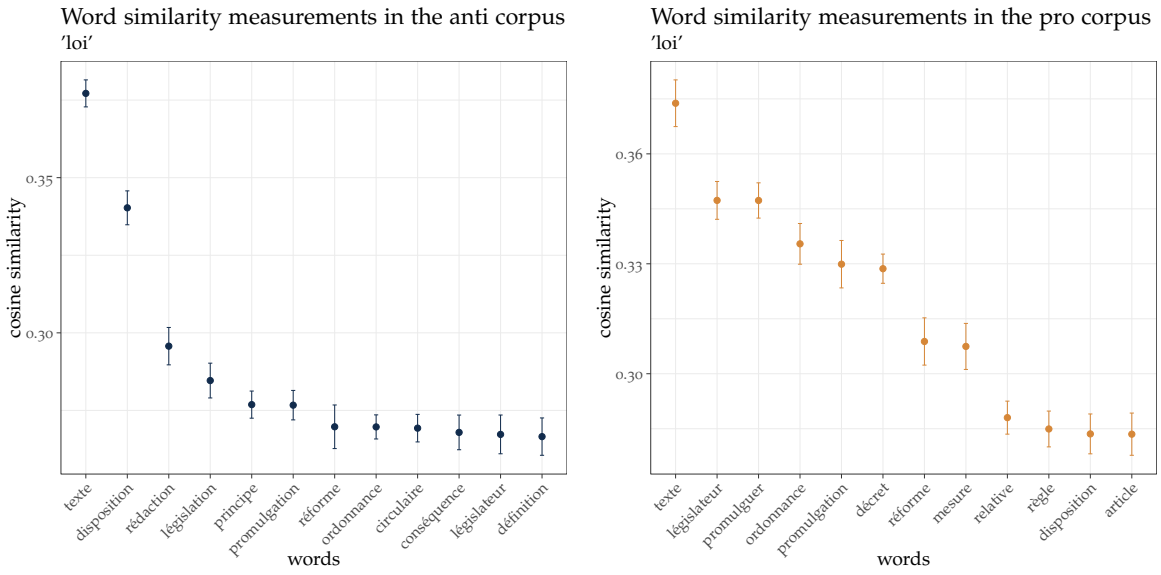


Figure 47: Comparison of closest semantic neighbours for *loi* across MPT models.

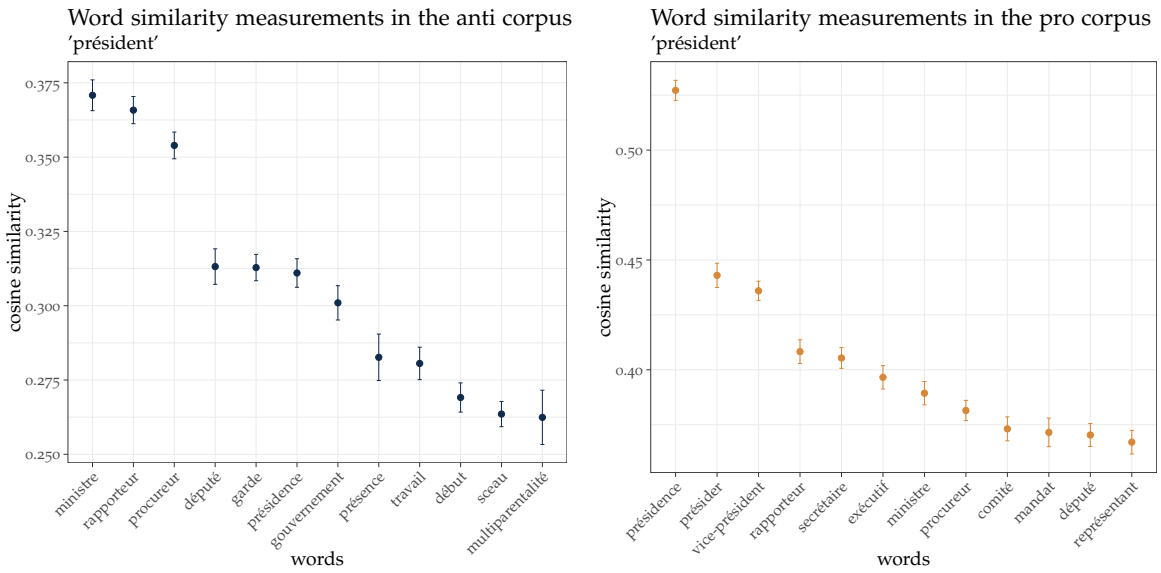


Figure 48: Comparison of closest semantic neighbours for *président* across MPT models.

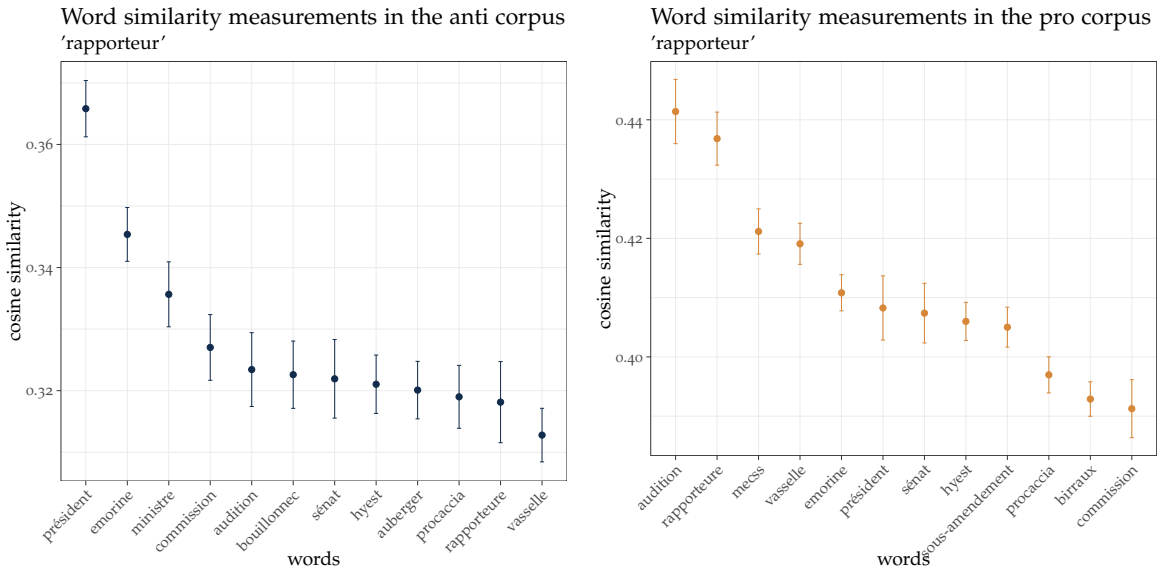


Figure 49: Comparison of closest semantic neighbours for *rapporteur* across MPT models.

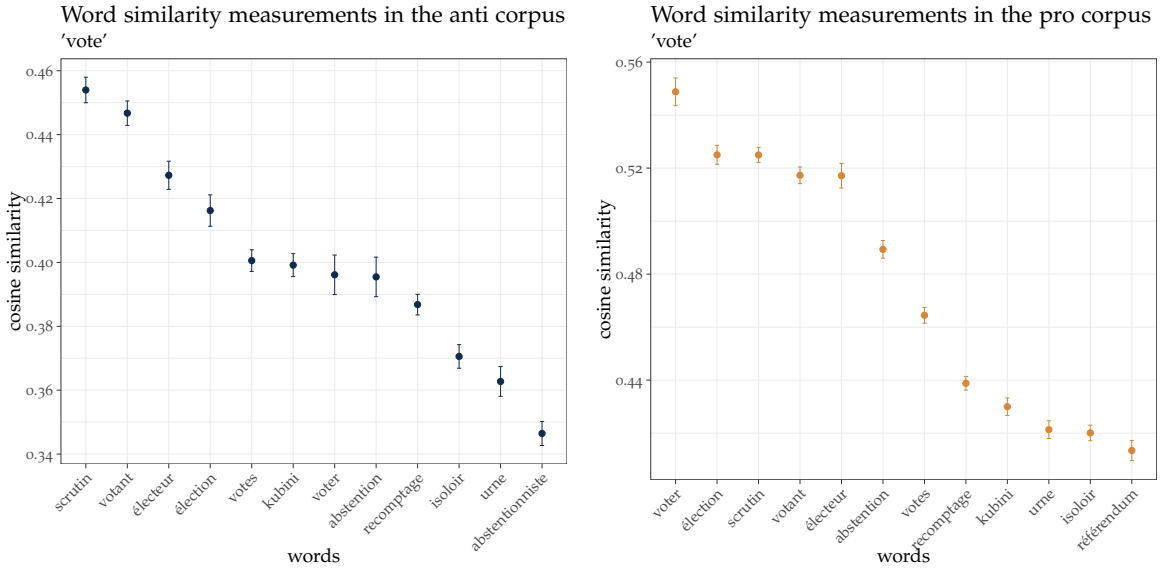


Figure 50: Comparison of closest semantic neighbours for *vote* across MPT models.

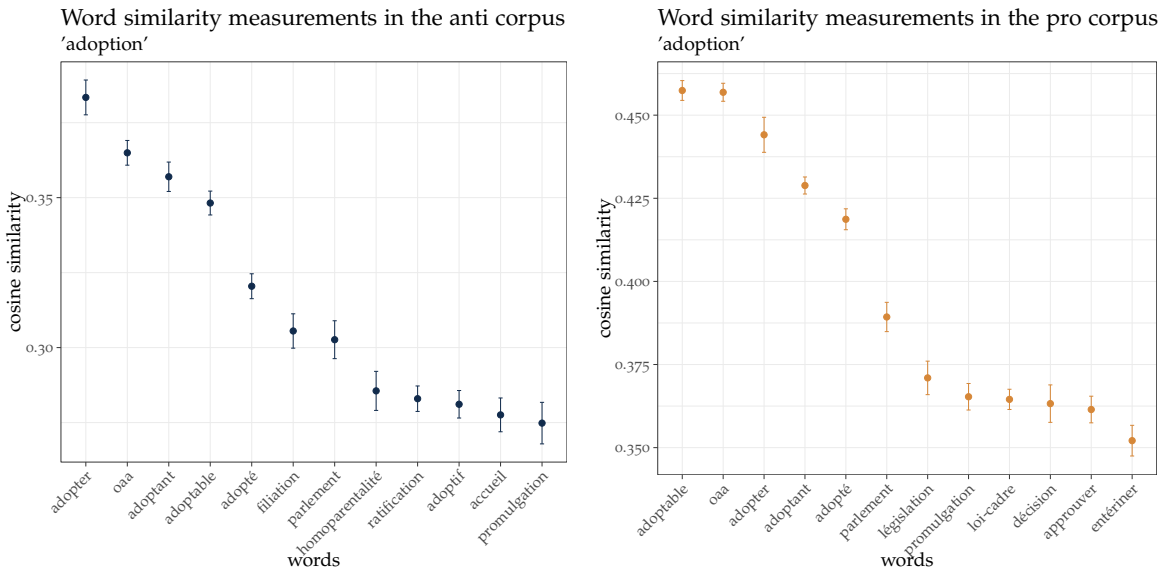


Figure 51: Comparison of closest semantic neighbours for *adoption* across PACS models.

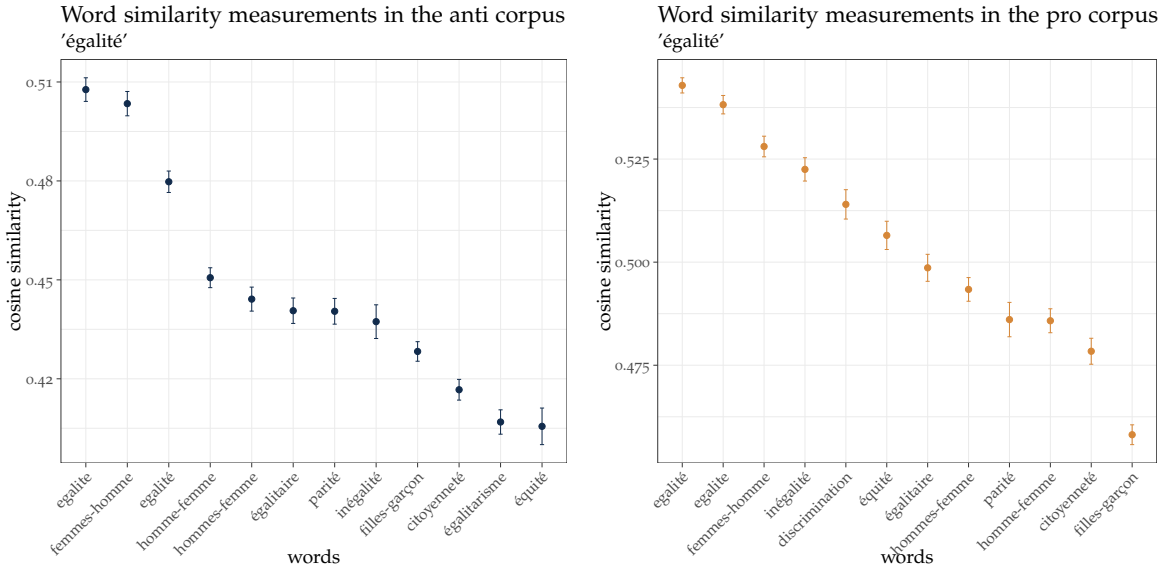


Figure 52: Comparison of closest semantic neighbours for *égalité* across PACS models.

D.2 PACS

D.2.1 *Test words*

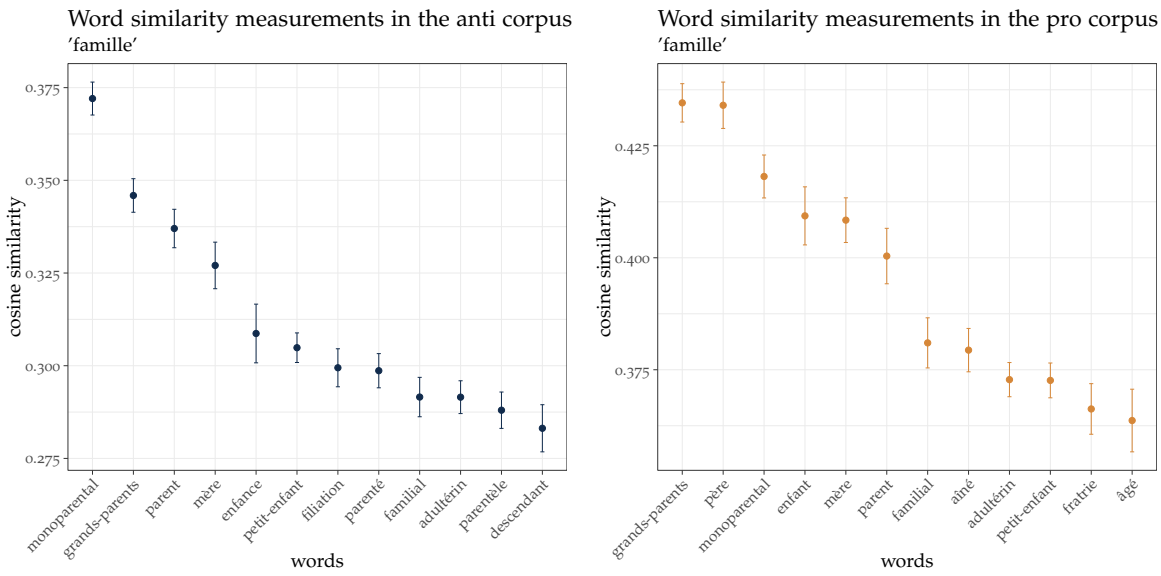


Figure 53: Comparison of closest semantic neighbours for *famille* across PACS models.

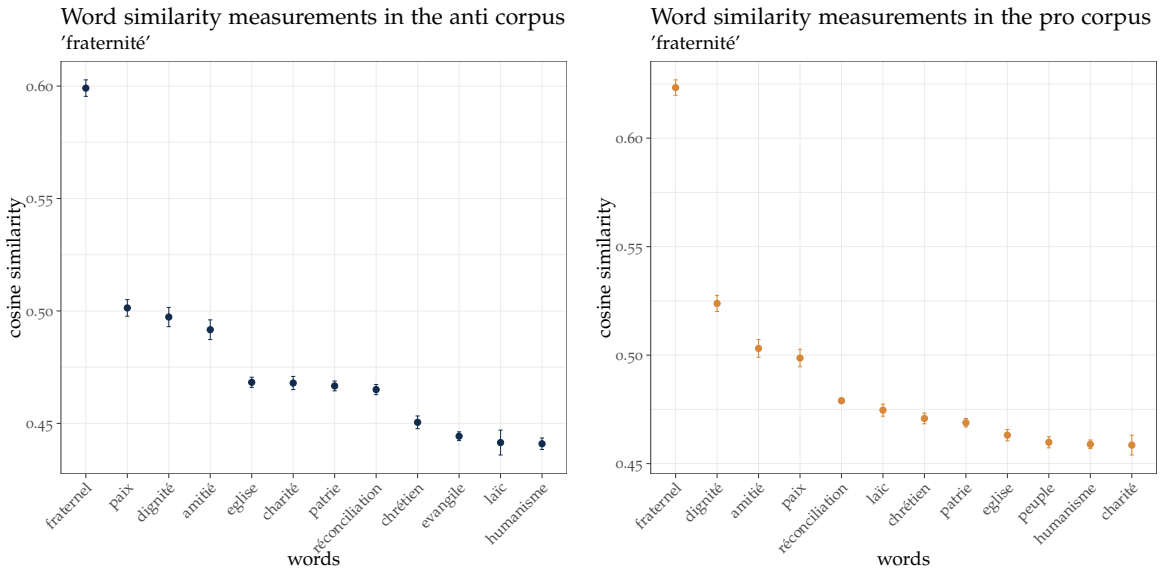


Figure 54: Comparison of closest semantic neighbours for *fraternité* across PACS models.

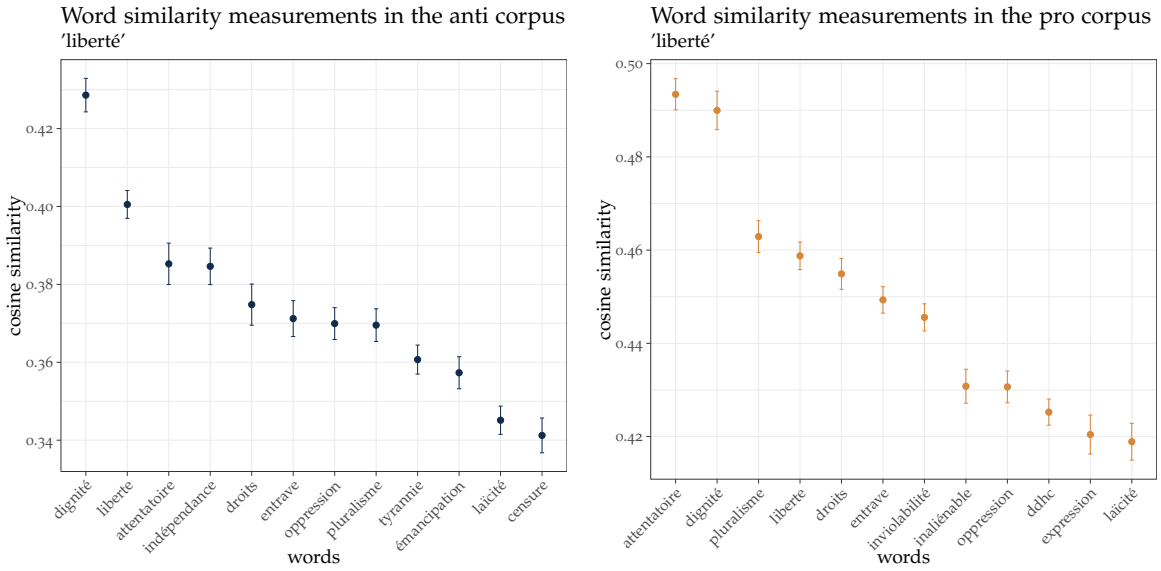


Figure 55: Comparison of closest semantic neighbours for *liberté* across PACS models.

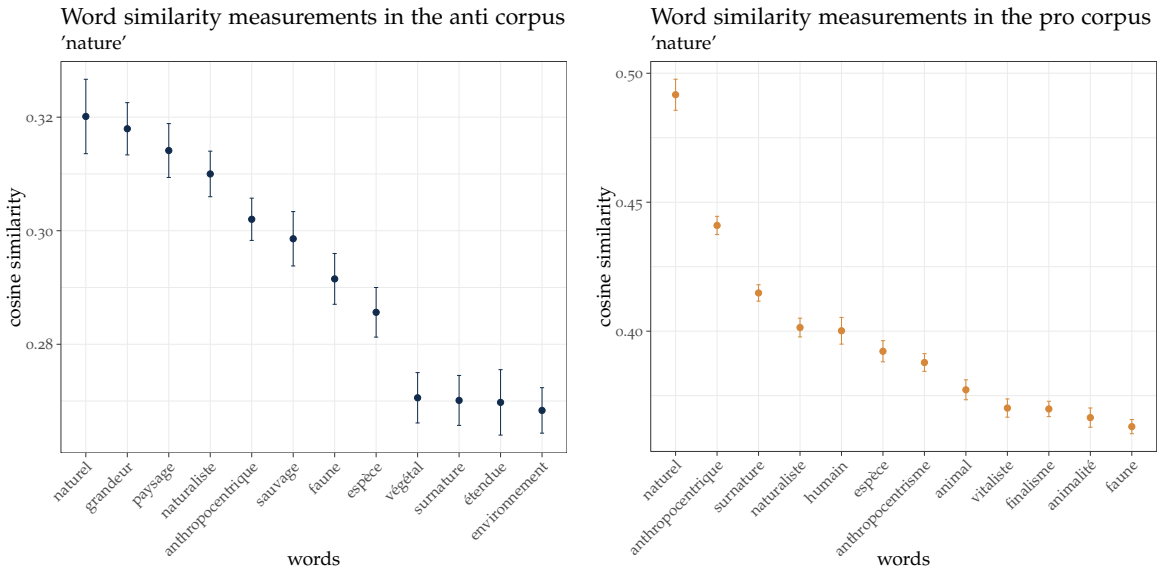


Figure 56: Comparison of closest semantic neighbours for *nature* across PACS models.

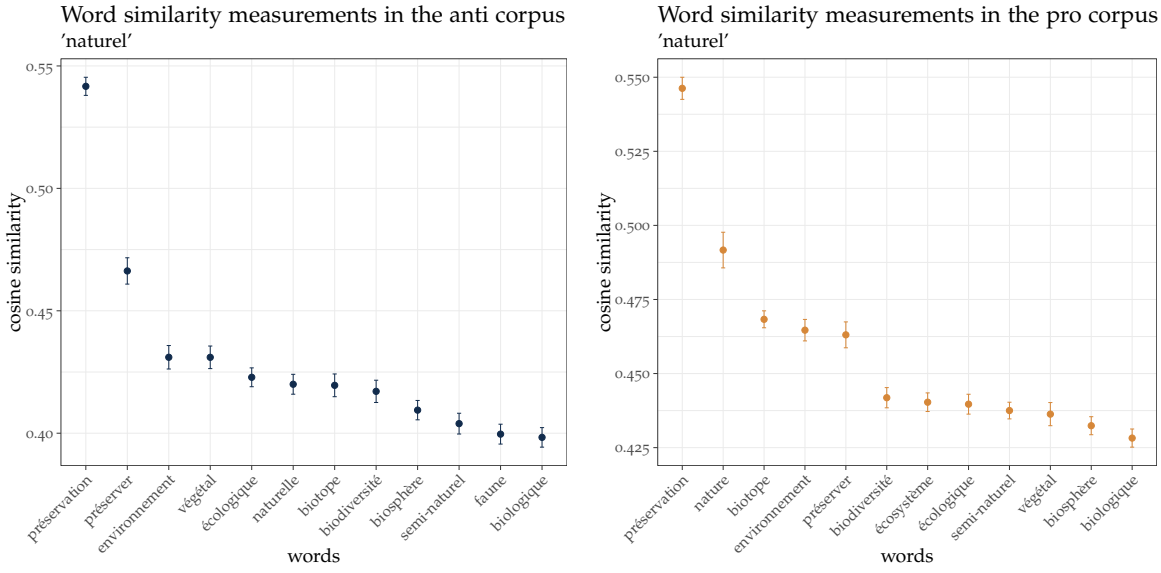


Figure 57: Comparison of closest semantic neighbours for *naturel* across PACS models.

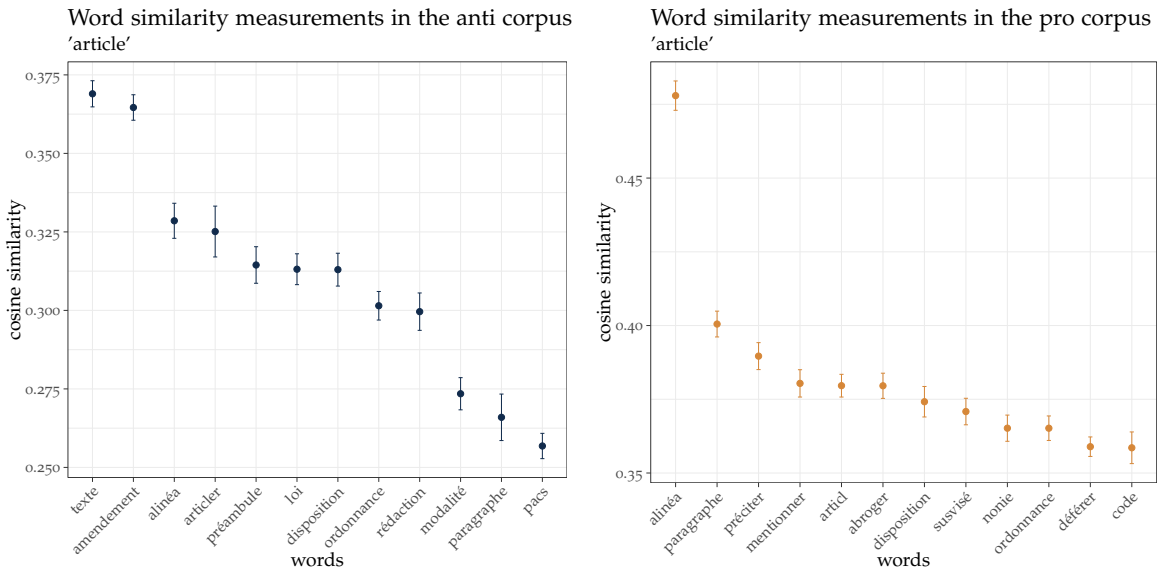


Figure 58: Comparison of closest semantic neighbours for *article* across PACS models.

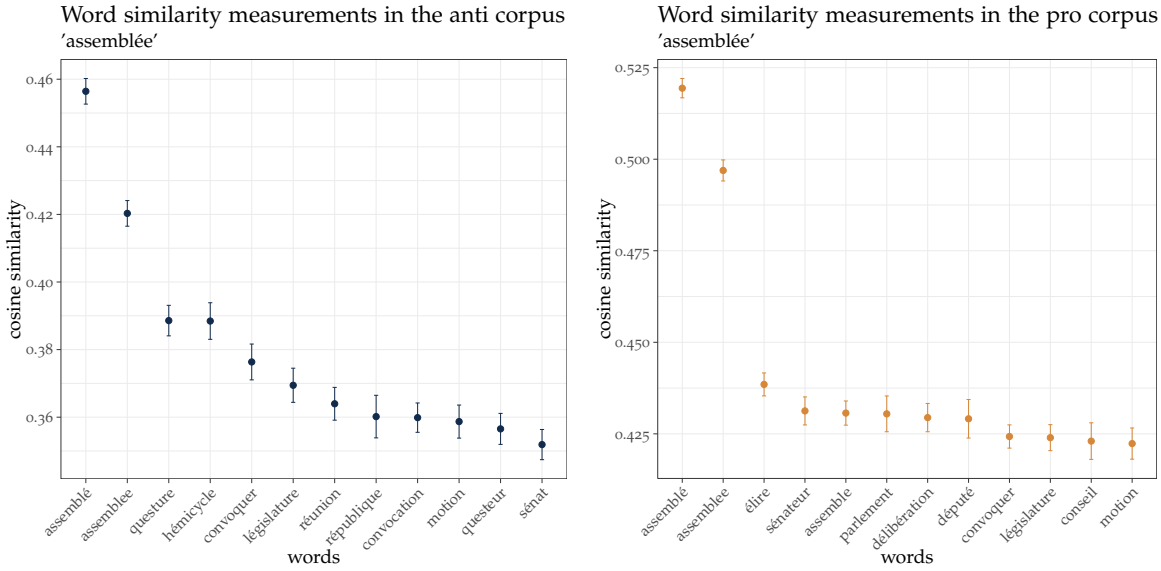


Figure 59: Comparison of closest semantic neighbours for *assemblée* across PACS models.

D.2.2 *Control words*

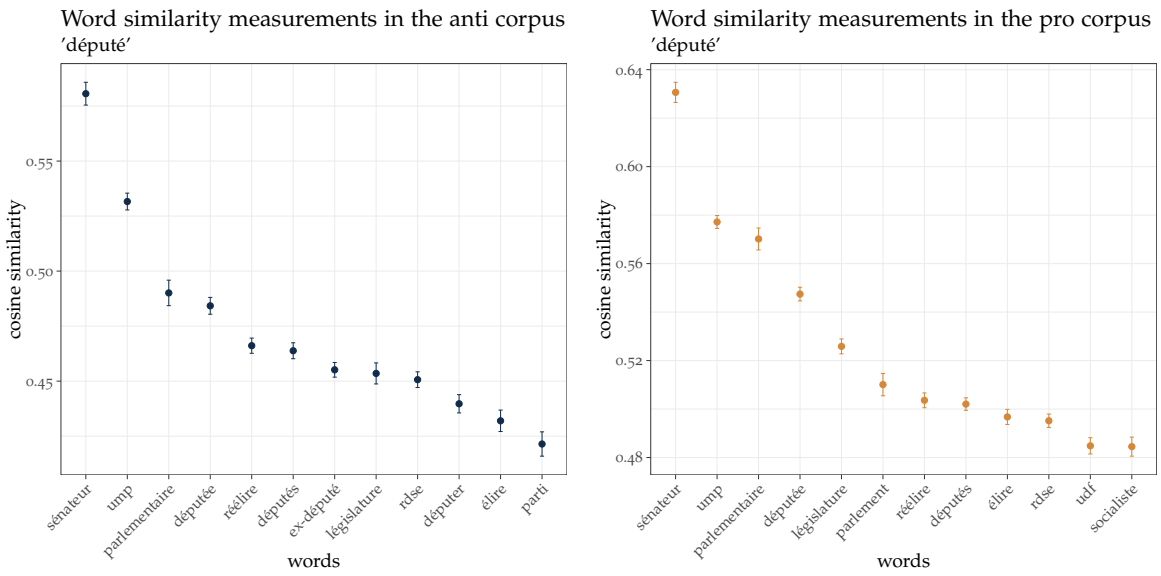


Figure 60: Comparison of closest semantic neighbours for *député* across PACS models.

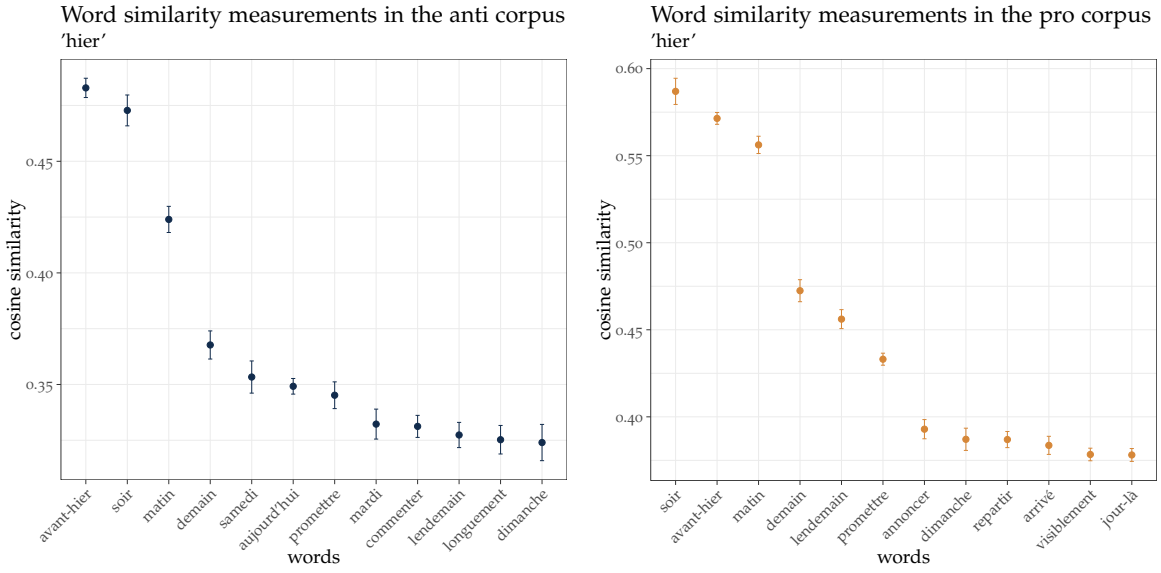


Figure 61: Comparison of closest semantic neighbours for *hier* across PACS models.

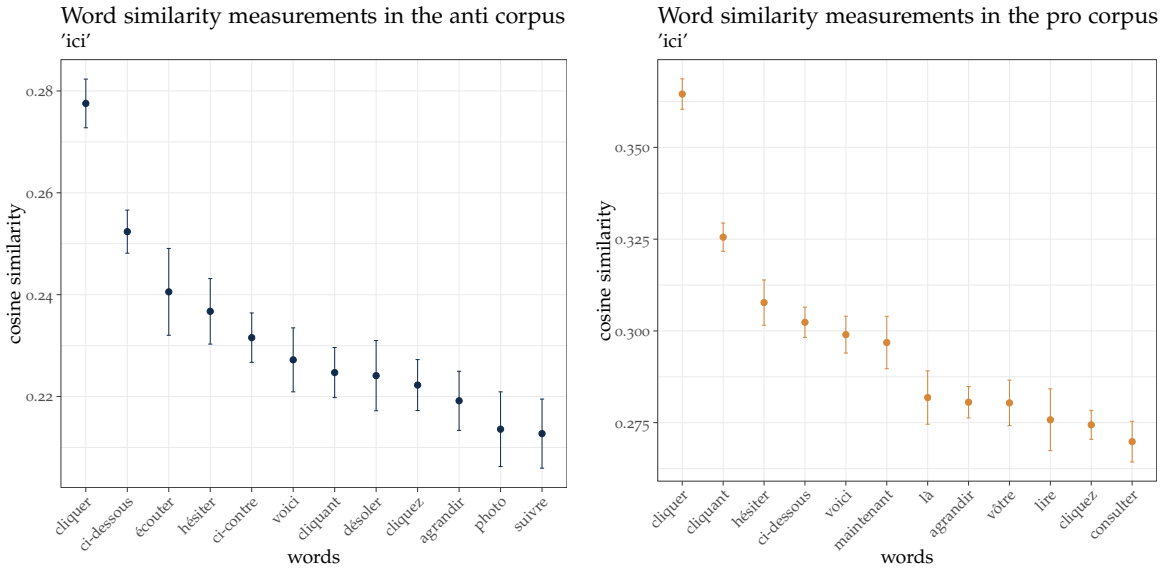


Figure 62: Comparison of closest semantic neighbours for *ici* across PACS models.

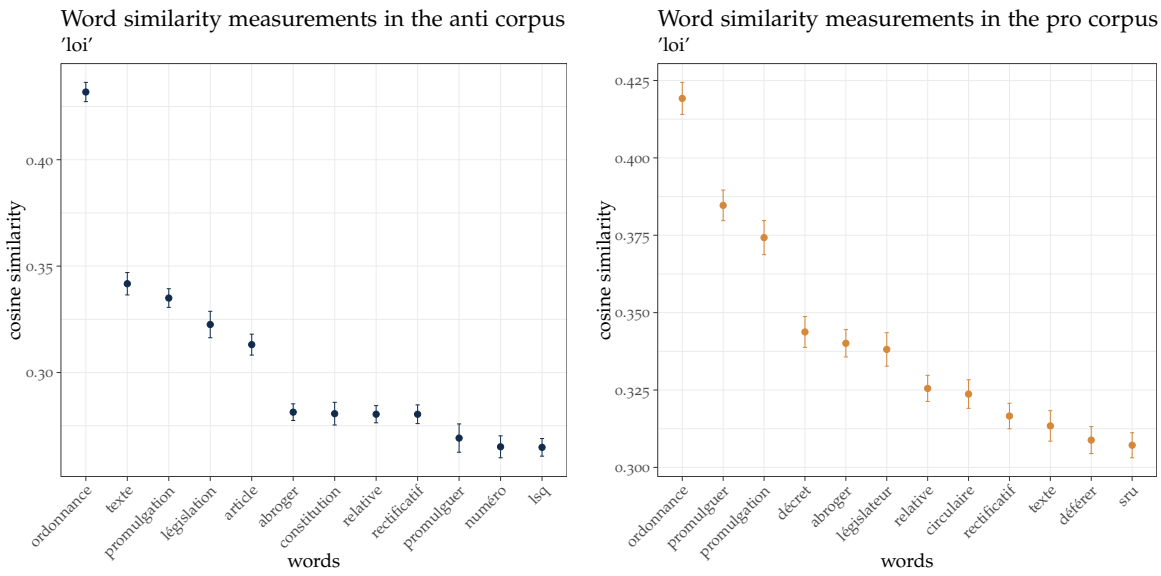


Figure 63: Comparison of closest semantic neighbours for *loi* across PACS models.

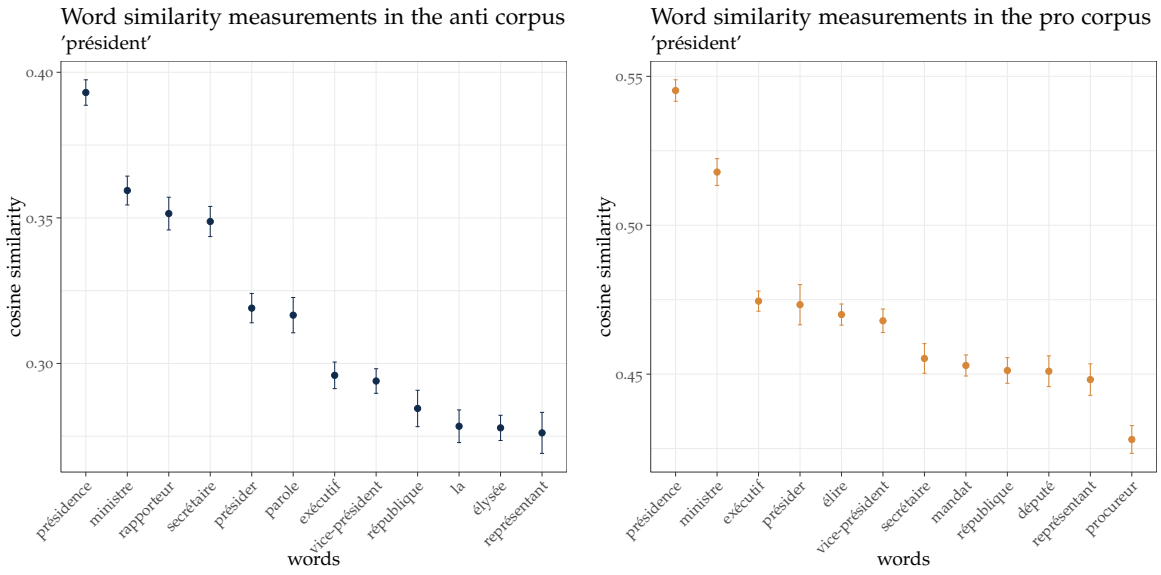


Figure 64: Comparison of closest semantic neighbours for *président* across PACS models.

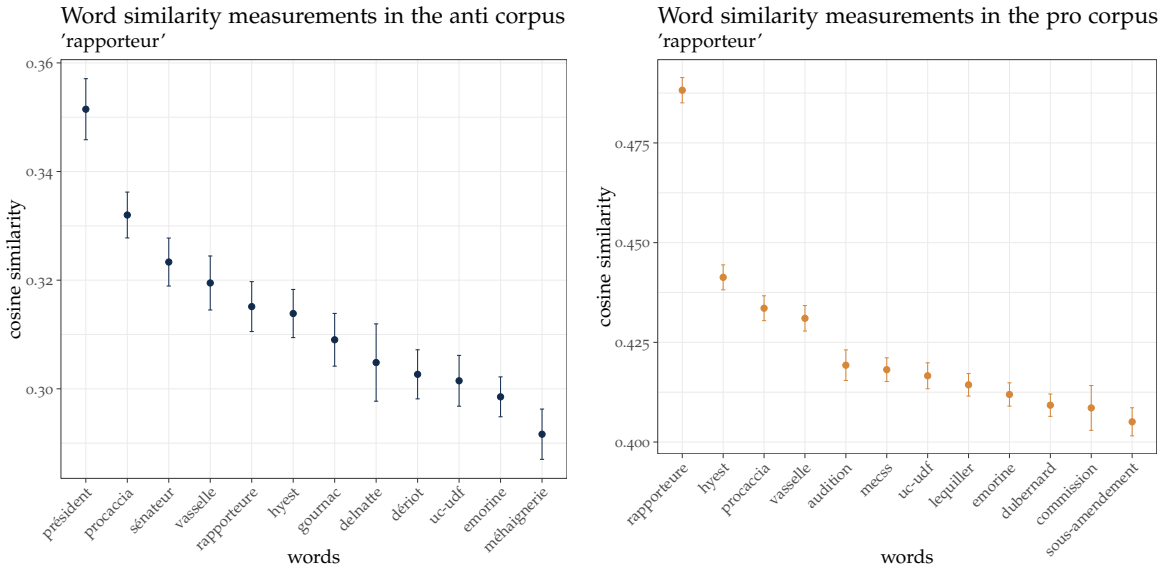


Figure 65: Comparison of closest semantic neighbours for *rapporteur* across PACS models.

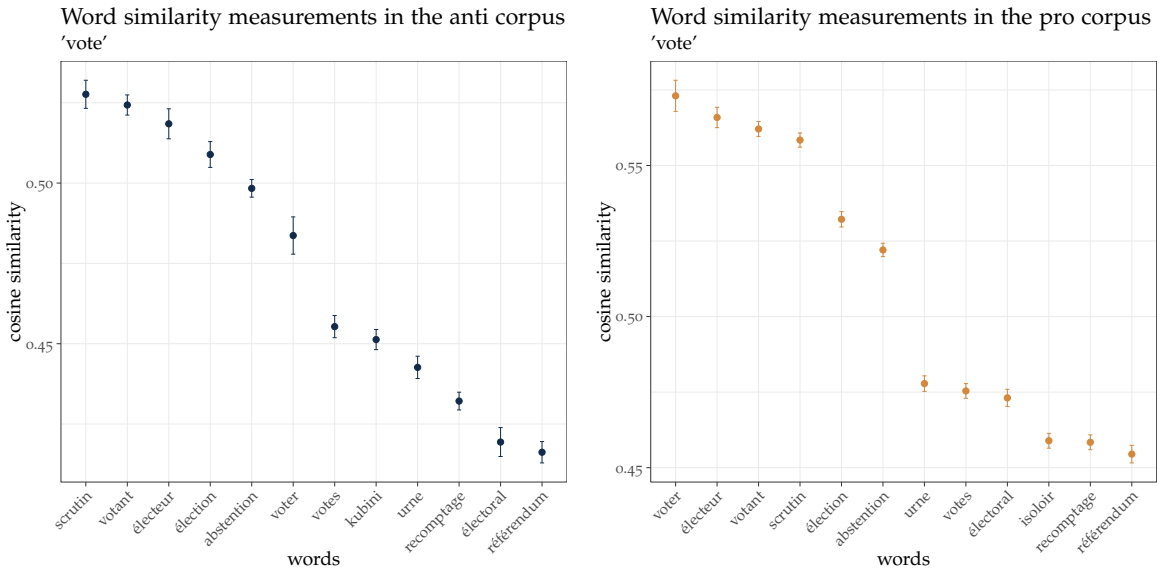


Figure 66: Comparison of closest semantic neighbours for *vote* across PACS models.