Um and Uh, and the expression of stance in conversational speech

Esther Le Grézause

Thèse de doctorat de l'Université Sorbonne Paris Cité et de l'Université de Washington et Préparée à l'université de Washington et à l'Université Paris Diderot

ECOLE DOCTORALE :
École Doctorale "Sciences du Langage" ED 132

DOCTORAT EN COTUTELLE INTERNATIONALE DE THESE

Um and *Uh,* and the Expression of Stance in Conversational Speech
Um et *Uh,* et l'expression de la prise de position dans le discours conversationnel

Par Esther Le Grézause

Thèse de doctorat de Linguistique

Dirigée par Nicolas Ballier et Richard Wright

Présentée et soutenue publiquement à l'Université de Washington
à Seattle, WA, le 23 Mai, 2017

JURY
Ballier, Nicolas, Co-directeur de thèse, professeur à l'Université Paris-Diderot, Sorbonne Paris Cité
Hanote, Sylvie, Rapporteur, professeure à l'Université de Poitiers
Herment, Sophie, Rapporteur, professeure à l'Université d'Aix-en-Provence
Ellen Kaisse, membre du comité de thèse, professeure à l'Université de Washington
Gina-Anne Levow, Présidente du jury, professeure à l'Université de Washington
Mari Ostendorf, membre du comité de thèse, professeure à l'Université de Washington
Richard Wright, Co-directeur de thèse, professeur à l'Université de Washington

Um and *Uh*, and the Expression of Stance in Conversational Speech

Esther Le Grézause

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Richard Wright, Chair

Nicolas Ballier

Gina-Anne Levow

Mari Ostendorf

Ellen Kaisse

Program Authorized to Offer Degree:
UW Department of Linguistics

University of Washington

Abstract

Um and *Uh*, and the Expression of Stance in Conversational Speech

Esther Le Grézause

Chair of the Supervisory Committee:
Professor Richard Wright
Department of Linguistics

Um and *uh* are some of the most frequent items in spoken American and British English (Biber et al., 1999). They have been traditionally treated as disfluencies but recent research has focused on their discursive functions and acoustic properties, and suggest that *um* and *uh* are not just filled pauses or random speech errors. At this stage, there is little agreement on whether they should be considered as by-products of the planning process (speech errors) or as pragmatic markers. In addition, most work on *um* and *uh* considers them to be the same variable, collapsing both into the same category.

The present work investigates the discursive and the acoustic properties of *um* and *uh* in spontaneous speech with the aim of finding out if they occur in systematic ways and if they correlate with specific variables. The analysis of *um* and *uh* is conducted on two corpora, ATAROS and Switchboard, to determine how the markers are used in different spontaneous speech activities. The Switchboard corpus consists of phone conversations between strangers, which allow us to study how speakers use *um* and *uh* in this context. It has different transcript versions (original and corrected), which allows us to test how transcribers perceive the two markers by aligning the original transcript with the corrected one. The ATAROS corpus consists of collaborative tasks between strangers and it is annotated for stance strength and polarity, which allows us to investigate how *um* and *uh* relate to stance. The term *stance* refers to subjective spoken attitudes toward something (Haddington, 2004). Stance *strength*

is the degree to which stance is expressed. Stance strength has four possible values : no stance, weak, moderate, and strong stance. Stance *polarity* is the direction of the expression of stance, and it can be positive, neutral, or negative.

The results of this study show that *um* and *uh* have different discursive cues, which correlate with variables such as speaker, speaker gender, speaker involvement, and naturalness of the conversation. *Um* and *uh* have different acoustic cues, which show some correlation with different degrees of stance and with stance polarity for different acoustic properties depending on the marker. The presence and the position of *um* and *uh* in utterances affect the likelihood of an utterance to be marked with a certain degree or polarity of stance. These findings are incorporated in a classification experiment, to test whether information pertaining to *um* and *uh* can be used to train a classifier to automatically label stance. The results of this experiment reveal that *um* and *uh* are valuable word unigram features and indicate that the position features and certain acoustic features increase the performance of the system in predicting stance. The results also indicate that *um* is a slightly more important word unigram feature than *uh*, that features pertaining to *um* are more informative in the prediction of binary stance, and that features relating to *uh* are more informative to predict three-way stance strength.

The findings confirm that *um* and *uh* are distinct entities. The discourse and acoustic features of *um* and *uh* are different. The marker *um* tends to vary to a greater extent than the marker *uh*. Transcribers perceive *um* more reliably than *uh*. *Um* and *uh* are relevant word unigram features. Features associated to *um* increase accuracy over those related to *uh* to predict binary stress, and features associated to *uh* increase accuracy over those associated to *um* to predict three-way stance strength. The work presented in this dissertation provides support to show *um* and *uh* are not just fillers or disfluencies, but rather that they have a wide range of uses, from fillers to pragmatic and stance markers.

Résumé de la thèse

Um et *uh* font partie des mots les plus communs en anglais parlé américain et britannique (Biber et al., 1999). Ces entités linguistiques ont souvent été traitées en tant que disfluence mais plusieurs études récentes se sont concentrées sur leur fonctions discursives et acoustiques, suggérant que *um* et *uh* ne sont pas juste des pauses pleines ou des erreurs de parole aléatoires. À ce stade, leur statut ne fait pas consensus. *Um* et *uh* sont généralement considérés comme des erreurs de production ou comme des marqueurs de discours. Ainsi, dans cette thèse, *um* et *uh* sont désignés par le terme "*marqueur*" (« *marker* ») afin d'inclure et de prendre en compte la majorité des rôles traditionnels qui leurs ont été attribués : pause pleine, disfluence, marqueur de discours, conjonction ou marqueurs de tour de parole. De plus, la majorité des études concernant *um* et *uh* les regroupe dans la même catégorie, malgré plusieurs analyses suggérant que l'un et l'autre ont des rôles différents.

Les objectifs de cette analyse sont d'explorer le rôle discursif et les propriétés acoustiques des deux marqueurs, de trouver s'ils sont utilisés de manière systématique, s'ils sont corrélés avec certaines variables spécifiques relevant du discours, s'ils sont liés à l'expression de la prise de position, et de confirmer que *um* et *uh* sont des entités distinctes en utilisant de nouvelles données et de nouvelles variables. Pour cela, *um* et *uh* sont analysés séparément et deux corpus, ATAROS et Switchboard, permettent de déterminer comment ils sont utilisés dans différentes tâches de communication.

Le corpus Switchboard est composé d'enregistrements d'entretiens téléphoniques entre deux inconnus conversant autour d'un sujet qui leur est imposé. Ce corpus comprend différentes versions de transcription qui permettent d'analyser la façon dont les transcripteurs perçoivent *um* et *uh* grâce à l'alignement des transcriptions originales avec les transcriptions

corrigées. Ce corpus permet donc d'étudier la façon dont les locuteurs utilisent *um* et *uh* dans ce contexte, ainsi que la manière dont les transcripteurs perçoivent ces deux entités. Le corpus ATAROS se compose d'enregistrements de tâches collaboratives entre deux inconnus et comprend des annotations sur le degré et la polarité de la prise de position, permettant ainsi l'analyse de *um* et *uh* en lien avec la prise de position. Le terme "stance" utilisé dans cette thèse pour désigner la prise de position est emprunté à Haddington (2004), et désigne l'attitude des locuteurs par rapport à ce qu'ils disent. Le degré de la prise de position se réfère à l'intensité avec laquelle le locuteur exprime ses propos. L'intensité correspond au degré de prise de position et peut prendre quatre valeurs : pas de prise de position, prise de position faible, modérée ou forte. La polarité de la prise de position correspond à son appréciation, qui peut être positive, neutre ou négative.

Cette thèse est composée de trois parties principales. La première partie aborde la mise en place des concepts et des données (chapitres 1 à 5). La deuxième partie porte sur les analyses discursives, acoustiques, et perceptives (chapitres 6 à 9). La dernière partie de la thèse propose une analyse de type apprentissage automatique qui rassemble les différents ordres de problématiques analysées (chapitres 10 et 11). Dans le détail, le chapitre 1 sert d'introduction à la thèse, pose les problématiques et les méthodes, remet en perspective les enjeux et annonce le plan suivi. Le chapitre 2 définit les principaux types de disfluences (cliniques et naturelles), résume les études principales conduites sur les disfluences, et présente les différents points de vue sur leur rôle dans le discours. Le chapitre 3 dresse l'état de la question sur le statut des deux pauses pleines (fillers) *um* et *uh* et montre comment plusieurs études récentes accréditent l'idée d'une différence pragmatique, voire fonctionnelle, entre ces deux "*fillers*", qu'il convient donc d'envisager comme des marqueurs. Le chapitre 4 revient sommairement sur le concept de "stance" (prise de position, évaluation), établit son acception dans cette thèse et dans l'annotation du corpus ATAROS, puis présente l'état de la question quant à la détection automatique de "*stance*" dans les corpus oraux. Le chapitre 5 caractérise

les deux corpus étudiés, ATAROS et Switchboard (SWB), et établit leurs contributions à l'analyse. Ce chapitre présente les méthodologies d'annotation des corpus, les deux versions de SWB, ainsi que la méthode suivie pour construire une interopérabilité entre ces deux corpus pour l'analyse de *um* et *uh*. Le chapitre 6 analyse la distribution et la durée des deux marqueurs dans SWB et ATAROS en fonction du genre des interlocuteurs, de l'authenticité de la conversation, et du nombre de conversations auxquelles les sujets participent. Ce chapitre montre que *um* et *uh* ont des durées et des distributions différentes et indique que les marqueurs ne sont pas utilisés au hasard. Le chapitre 7 se penche sur la production de *um* et *uh* dans SWB, et sur la perception des deux marqueurs en comparant les deux versions des transcriptions du corpus. Trois types d'erreurs sont dégagées des comparaisons entre les deux versions de transcription : les substitutions, les oublis et les mots inventés par les transcripteurs alors que les locuteurs ne les ont pas produits ("*hallucinations*"). Les principaux résultats montrent que *um* et *uh* sont plus souvent oubliés que d'autres mots fréquents tels que les mots outils, et que les transcripteurs de SWB font plus d'erreurs sur *uh* que sur *um*, suggérant que *um* joue un rôle discursif plus important que *uh*. Le chapitre 8 interroge la relation entre la prise de position ("*stance*") d'une unité de parole ("*spurt")* et la présence et la position des marqueurs dans cette unité, et révèle que ces deux dimensions sont dépendantes, et que les résultats sont différents pour les deux marqueurs. Le chapitre 9 évalue la relation entre la prise de position d'une unité de parole et la réalisation acoustique de la voyelle des marqueurs, comparée à la même voyelle dans d'autres mots monosyllabiques. Les résultats indiquent que les valeurs de "*stance*" affectent avec différents degrés la réalisation acoustique des marqueurs. Le chapitre 10 se fonde sur les résultats des expériences précédentes (chapitre 6, 8 et 9) pour plusieurs tâches de classification qui testent les traits ("*features*") les plus importants pour prédire automatiquement les valeurs de "*stance*" en fonction des paramètres correspondants à *um* et *uh* (traits lexicaux, positionnels et acoustiques). Ces expériences montrent que les traits pertinents de ces marqueurs affectent la performance du système

et que les meilleurs résultats de la classification sont obtenus lorsque les traits lexicaux *um* et *uh* sont présents, et lorsque leur position est prise en compte. Les résultats indiquent également que les algorithmes qui prennent en compte les différentes propriétés acoustiques améliorent les scores de prédiction. Le chapitre 11 conclut la thèse en résumant les résultats des chapitres 6 à 10, en soulignant les impacts de cette recherche, et en indiquant les pistes de recherche futures.

Pour résumer, cette étude montre que *um* et *uh* sont utilisés de manière différente dans les discours oraux. Chaque marqueur est corrélé à des variables telles que le locuteur, le genre du locuteur, l'investissement du locuteur (selon la tâche du corpus ATAROS) et l'authenticité ("*naturalness*") de la conversation annotée par les transcripteurs de SWB. La présence et la position des deux marqueurs affectent la probabilité qu'une proposition soit affectée d'un certain degré de polarité ou d'un certain degré de prise de position, et *um* et *uh* ont alors des réalisations acoustiques différentes selon les valeurs de prise de position. Ces résultats sont intégrés et exploités dans plusieurs expériences de classification automatique afin de tester si l'information concernant *um* et *uh* peut être utilisée afin d'optimiser et d'améliorer la reconnaissance automatique de prise de position dans les conversations. Les conclusions de ces expériences révèlent que *um* et *uh* sont des traits lexicaux importants et indiquent que leurs propriétés acoustiques et, plus encore, leur présence et leur position dans une proposition sont des traits pertinents pour la classification automatique de la prise de position. Pour conclure, le travail conduit dans cette thèse étaye les résultats des travaux précédents qui montrent que le rôle de *um* et *uh* ne se limite pas au statut de pause pleine ou de disfluence, et confirme une utilisation qui va de l'erreur de parole au marqueur de discours, en passant par le marquage du degré et de la polarité de la prise de position.

Cette étude adopte une approche de type multidimensionnel afin de mieux comprendre les usages et les fonctions de *um* et *uh* dans les conversations spontanées en anglais américain. Cependant, étant donné le nombre de dimensions considérées, il reste encore plusieurs angles

à explorer. Les prochaines étapes de cette analyse se concentreront sur le contexte acoustique des marqueurs (fluctuations de l'intonation par rapport à l'environnement immédiat autour des marqueurs, le troisième formant et le type de phonation), le contexte et la complexité syntaxique autour des marqueurs, et une étude plus approfondie sur la perception de *um* et *uh* en fonction de divers environnements dans lesquels les deux marqueurs sont produits.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

ATAROS: "Automatic Tagging and Recognition of Stance", funded by NSF IIS #1351034 Audio corpus of collaborative tasks between groups of two speakers with high quality recordings, designed to elicit different degrees of involvement and stance, to look at the acoustic signal of stance-taking (Chapter 5). More information on the corpus can be found on the Linguistic Phonetic Lab website : `http://depts.washington.edu/phonlab/projects.htm`

CLASSIFIER: a machine learning tool that takes data items and places them into one of k classes (Chapter 10).

DISFLUENCY: common phenomenon in spontaneous speech that interrupts the normal flow, commonly divided into two main categories : clinical and normally-occurring (Chapter 2)

DISFLUENCY - CLINICAL: occurs in the speech of people who have a disorder that might be directly linked to speech, or that causes speech impediments. Common distinctions differentiate between stutter-like and non stutter-like disfluencies (section 2.3)

DISFLUENCY - NORMALLY OCCURRING: naturally occurring phenomenon in spontaneous speech that interrupts the normal flow. Disfluency types range from filled pauses, repetitions, false starts, substitutions, insertions, and speech errors (Shriberg, 2001) (section 2.4)

DYAD: a group of two speakers

K-MEANS CLUSTERING: used for cluster analysis and separates samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares (Chapter 10).

MARKER: term used in this document to reference *um* and *uh* without referring to their function such as *filled pause*, *disfluency*, or *discourse marker* (section 1.1)

SLASH UNIT: sentence-like chunks of speech used to segment the Switchboard corpus transcriptions (section 5.4.3)

SPURT: corresponds to a sequence of speech between pauses greater than 500ms by one speaker (Shriberg et al., 2001a), comparable to discourse structure based unit, related to intonation, and used as the unit for stance annotation (section 5.3.5)

STANCE: overt communicative act that uses language, expressed by a stancetaker (Du Bois, 2007). Refers to "the speakers' subjective attitudes toward something" (Haddington, 2004, p. 101) (Chapter 4).

STANCE POLARITY: stance annotation (Freeman, 2015) with 3 levels that correspond to the polarity of the speaker's attitude : positive, neutral, and negative (sections 4.2 and 5.3.4)

STANCE STRENGTH: stance annotation (Freeman, 2015) with 4 levels that correspond to the degree of the speaker's attitude : no stance, weak, moderate, and strong stance (sections 4.2 and 5.3.4). Other stance strength distinctions include binary stance : no stance vs. any degree of stance ; and three-way stance : no stance, weak stance, and moderate-to-strong stance

STANCE-TAKING: the act of expressing a stance act (Chapter 4)

SVM - SUPPORT VECTOR MACHINE: supervised learning method used for classification, regression and outliers detection (Chapter 10).

SWITCHBOARD: corpus of spontaneous conversational speech. Multi-speaker database of telephone bandwidth speech, with time aligned word transcriptions and a variety of annotations depending on the version or the subset of the corpus (Godfrey et al., 1992) (Chapter 5)

TOKEN FREQUENCY: relative frequency of occurrence of a word within a specific dataset or a corpus

ACKNOWLEDGMENTS

This dissertation was realized as a co-sponsorship between the University of Washington and the University of Paris Diderot - Sorbonne Paris Cité, with my advisers Dr. Richard Wright and Dr. Nicolas Ballier.

First and foremost I would like to express my sincere gratitude to my two advisers, Dr. Richard Wright and Dr. Nicolas Ballier, for your continuous support throughout my time in graduate school, for your kindness, and for sharing your immense knowledge with me. I would also like to thank you both for making the joint-thesis sponsorship happening, and for collaborating. Thank you Dr. Richard Wright for encouraging me to come to the University of Washington, for supporting all my research interests, for your guidance, for your support, and for bringing me on an incredible project that relates to my research. Thank you Dr. Nicolas Ballier for your ongoing presence on the other side of the globe, for your interests and investments in my research, for all your detailed feedback, and for initiating the joint-thesis sponsorship. I could not have asked for better mentors, and I would like you to know that I am grateful for everything you did, for all your time, and for your encouragements.

I would also like to express special thanks my committee members at the University of Washington, Dr. Gina-Anne Levow, Dr. Mari Ostendorf, and Dr. Ellen Kaisse, for your time, for your guidance, and for helping me improving my work and research. I am grateful I was able to work with you and to learn from you.

Thanks to the *Prosodic Anomalies* group, Dr. Richard Wright, Victoria Zayats, Trang Tran, and Dr. Mari Ostendorf, who helped me collect portions of the data, and who made an entire chapter of this dissertation possible. Special thanks to Dr. Mari Ostendorf for letting me collaborate on this very interesting project, and for her guidance throughout the project.

Thank you to the two external reviewers (pré-rapporteur), Sophie Herment, Professor at the University of Aix-en-Provence, and Sylvie Hanote, Professor at the University of Poitiers, who kindly accepted to read and report on my dissertation in short delays.

I would also like to thank my friend and colleague Laura Panfili, who shared useful LaTeXtips with me, who brought me food, and who proof-read several chapters of my dissertation. On the same note, thanks to my friend Daniel Shapero, who proof-read a chapter of my dissertation, and who encouraged me. Thanks to my friend and colleague Marina Oganyan who joined me to work on our respective dissertations, and for lightening our spirits. Finally, thanks to my friend Andrea Kahn who shared very useful machine learning tips with me.

Thanks to Valerie Freeman for her ground work on the ATAROS corpus, and for sharing LaTeXadvice. Thanks to the members of the Phonetics lab for their feedback and for sharing resources, especially John Riebold. Sincere thanks to the Department of Linguistics at the University of Washington, who takes great care of their students, and special thanks to Mike Furr and Joyce Parvi for their encouragements and for facilitating all administrative processes. I would also like to thank the Department of Etudes Anglophones at the University of Paris Diderot - Sorbonne Paris Cité for collaborating on this joint-thesis and for being accommodating.

I am grateful for my parents, Elisabeth and Francis, who encouraged me to study and to travel, and who gave me all the tools to succeed.

Last but not least, thanks to my dear husband, Andrew Baer, for his boundless support, kindness, and patience, for answering all my grammar questions, for proof-reading part of my work, and most importantly, for making me happy at all times.

Portions of this work were supported by National Science Foundation grant IIS-1617176, the University of Washington Department of Linguistics, and the University of Washington Department of Linguistics Excellence in Linguistic Research Graduate Award. Portions have

been presented at conferences and portions will be submitted for publication.

Chapter 1
INTRODUCTION

Spontaneous and conversational speech are characterized by the common presence of naturally-occurring disfluencies, such as repetitions, false starts, or filled pauses. Studies on conversational American English report that about 6% of words are disfluent (Fox Tree, 1995; Shriberg, 1994). Due to increasing interests in spontaneous speech, disfluencies have received attention in fields such as psycholinguistics, discourse analysis, and natural language processing. They are often characterized as natural by-product of speech planning, cognitive load, and turn-taking. Since the term *disfluency* encompasses a wide range of topics and dimensions, this dissertation focuses on two disfluencies : *um* and *uh*.

The goals of this dissertation are to explore how the two markers *um* and *uh* pattern depending on various discourse variables, whether they are the same entity, and how they correlate with the speaker's attitude (i.e., stance), in order to better understand how they function in spontaneous and conversational speech.

1.1 Um and uh

Um and *uh* are some of the most frequently occurring items in American and British spoken English (Biber et al., 1999; Tottie, 2015a). In this dissertation I am going to refer to *um* and *uh* as *markers* because this term encompasses several of their most traditional roles such as *filled pauses* or *disfluencies*, as well as other roles such as *discourse markers* and *speech act markers*. In this study, I use the terms *role* and *function* interchangeably to refer to the semantico-discursive properties of the two markers, and not just their syntactic properties often denoted by the term *function*, or their purely semantic characteristics often

conveyed by the term *role*.

Um and *uh* have been traditionally treated as disfluencies along the lines of repetitions, false starts, hesitations and pauses. Recent research (Clark and Fox Tree, 2002; Norrick, 2015; Tottie, 2014) however has focused on their status, properties, discourse functions, and environments. Various studies report different functions for the two markers, ranging from filled pauses, backchannels, interjections, to discourse markers. For instance, studies show that listeners use *um* and *uh* as cues for processing difficulty and online accessibility of referents (Arnold and Tanenhaus, 2011; Kidd et al., 2011; Watanabe et al., 2008). Studies also show that children start paying attention to the two markers around the age of two (Kidd et al., 2011). However, in subsequent processes of speech processing, several speech understanding systems filter disfluencies with the goal of improving system performance, despite the fact that several studies show that removing disfluencies increases the perplexity of the surrounding words. Moreover, we know that filled pauses like *um* and *uh* contain discourse and prosodic information such as marking linguistic units and restart boundaries or indicating hesitation when a speaker holds the floor (Siu and Ostendorf, 1996; Siu et al., 2000; Stolcke and Shriberg, 1996). Studies also show that *um* and *uh* have different acoustic properties in terms of pitch and vowel duration (Shriberg, 2001), and that they tend to occur sentence-initially (Clark and Fox Tree, 2002; Shriberg, 2001). In addition, several studies, including this study, show that the two markers occur in different environments depending on various linguistic and external factors, and argue that *um* and *uh* have different functions in speech (Clark and Fox Tree, 2002; Gorman et al., 2016; Irvine et al., 2015).

Such findings indicate that *um* and *uh* are not random errors, but rather that they have a function in discourse, and that they are part of the linguistic signal. The goals of this study are to explore the distribution, the characteristics, and the salience of the two markers; whether they relate to attitude marking (i.e., stance), and to confirm using new data and variables that *um* and *uh* are different entities.

1.2 Stance

Stance is an overt communicative act that takes place in language. It is expressed by a stancetaker, about the object of the stance act (Du Bois, 2007). Generally speaking, stance refers to "the speakers' subjective attitudes toward something" (Haddington, 2004, p. 101) and involves sociocultural, dialogical, linguistic, and intersubjective dimensions (Du Bois, 2007).

Stance is often referred to as *sentiment*, *emotion*, *subjectivity*, or *private state* (Quirk, 1985; Wilson and Wiebe, 2005). Recent studies show that stance in spontaneous spoken speech is marked by several elements not present in text, and that inter-annotator agreement on opinion categories is higher when transcribers have access to audio data (Freeman, 2014; Freeman et al., 2015b; Somasundaran et al., 2006). Other findings show *um* is an informative word unigram feature, and adding prosodic and speaking style features to the lexical features increases the system's accuracy when punctuation features are excluded (Levow et al., 2014).

These studies indicate the importance of taking acoustic cues and lexical features, such as *um*, into account. One of the main goals of this dissertation is therefore to investigate how different properties of *um* and *uh* relate to stance marking, and whether taking them into account improves the accuracy of automatic stance categorization.

1.3 Cross corpora study

I use two corpora in this dissertation, ATAROS and two versions of Switchboard. Both corpora allow an examination of how *um* and *uh* behave in conversational interactions consisting of different tasks, but also answer slightly different questions. The ATAROS (Automatic Tagging and Recognition of Stance) corpus is an audio corpus of collaborative tasks between groups of two speakers, designed to look at the acoustic signal of stance-taking (Freeman et al., 2014b). The Switchboard corpus is a multi-speaker database of telephone bandwidth speech (Godfrey et al., 1992), with a variety of annotations depending on the version or the subset of the corpus (Calhoun et al., 2010; Hamaker et al., 1998).

While having corpora with different speech tasks allows different questions to be answered, it also raises issues on cross-corpora studies, such as how to conduct a study on several datasets with different annotation guidelines, and how to compare findings. This raises several challenges moving forward. For example, having overarching guidelines for disfluency, partial word transcription, speech segmentation, pause annotation, or even recording techniques would provide benefits to research and to the interpretation of findings. Furthermore, cross-corpora validation of findings increases the robustness and the generalization of the findings.

1.4 Study goals and contributions

Relatively little work has been done on the acoustic and prosodic characteristics of stance in spontaneous speech, especially compared to text. Furthermore, very little to no attention has been given to the role of *um* and *uh* in signaling stance-taking in spontaneous speech. Moreover, many studies group the two markers under the same label, and very few focus on the differences between *um* and *uh*. Finally, there are no studies on the salience of the two markers compared to other words and to each other. This dissertation addresses these issues and gaps in the literature, by providing a systematic, holistic, quantitative, and multidimensional analysis of *um* and *uh* in conversational speech. The linguistic strata of this study encompass multiple dimensions ranging from token frequency, perception of the markers, to the acoustic properties and the discourse functions of *um* and *uh*. Note that this study exclusively relies on statistical analyses of the two markers ; it is strictly quantitative and does not incorporate a qualitative analysis of *um* and *uh*.

Preliminary studies show that the two markers have different distributions, which would indicate that they are not functionally the same. I also hypothesize that *um* is more salient than *uh*, which would suggest that *um* has more important discourse functions than *uh*. Based on the assumption that *um* and *uh* have functions in discourse, and based on the fact that stance is characterized by lexical and acoustic features, I predict that several characteristics of *um* and *uh* are robustly correlated with stance. This would suggest that *um* and *uh* can be used as stance markers, and therefore carry information about stance, which could then

be used for automatic stance recognition.

1.5 Study structure

Chapter 2, 3, 4 and 5 make up the background sections of this dissertation. Chapter 2 defines disfluencies, whether naturally-occurring or clinical, and reviews the different types of disfluencies, the factors that affect their production, and general findings on the topic. Chapter 3 is an exhaustive overview of the two markers *um* and *uh*. This chapter reviews a wide range of analyses, from online comprehension, production, to speech technology, and points out issues with existing studies. Stance is defined and reviewed in Chapter 4, and Chapter 5 describes in details the two corpora used in this study. The rest of the dissertation is organized into five experiments. The first experiment (Chapter 6) analyzes the distribution of the two markers in ATAROS and in the Mississippi State version of Switchboard, to determine if they are used interchangeably and in systematic ways across different speech tasks. In the next experiment (Chapter 7) I explore the factors that influence the production and the transcription errors made on *um* and *uh* in the Switchboard corpus, to find out they differ from other words and from each other. The next three experiments focus on the relationship between the two markers and several dimensions of stance marking (i.e., degree and polarity) in the ATAROS corpus. Chapter 8 investigates whether the presence and the position of *um* and *uh* in a speech unit affect the probability of the stance marking. In Chapter 9 I focus on the acoustic properties of the vowel /ʌ/ (AH1) in *um* and *uh*, to find out if stance marking affects the acoustic realization of the markers, and whether the vowel in the two markers behave differently from other monosyllabic words. In the last experiment of this dissertation (Chapter 10) I explore the relative contribution of *um* and *uh* to predicting stance strength and polarity by implementing classifiers that build on findings from this dissertation. Finally, the last chapter of this dissertation (Chapter 11) concludes this study by summarizing each experiment and their main findings, and states the contributions of this work as well as its future directions.

Chapter 2
DEFINING DISFLUENCIES

2.1 Introduction

The goals of this chapter are to define *disfluencies* and to explain why I am interested in studying them in spontaneous speech. I first provide a general introduction of the topic before going into different categories of disfluencies such as clinical and non-clinical disfluencies. I then summarize relevant work in the literature that focuses on any type of disfluencies, clinical and non-clinical, and their properties in speech. Finally, I talk about the challenges that disfluencies present for speech technology, especially for spontaneous speech.

2.2 General introduction on disfluencies

Spontaneous speech contains high rates of disfluencies (e.g. filled pauses, repetitions, false starts, etc.). About 6% of words are disfluent in conversational American English (Fox Tree, 1995; Shriberg, 1994, 1999). Disfluencies have received attention in the fields of psycholinguistics, discourse analysis, and natural language processing due to an increasing interest in spontaneous speech. A wide range of studies have looked at disfluencies. Some studies have worked to define what disfluencies are (Shriberg, 1994). Other studies have looked at their acoustic properties (Shriberg and Lickley, 1993; Shriberg, 1994, 1995, 1999, 2001), or automatic detection and modeling of speech disfluencies (Adda-Decker et al., 2004; Liu et al., 2006b,a; Marin and Ostendorf, 2014; Snover et al., 2004; Stouten et al., 2006; Wang et al., 2013; Zayats et al., 2014). Finally, the function of disfluencies in speech communication (Clark and Fox Tree, 2002; Fox Tree, 1995, 1997, 2001; O'Connell and Kowal, 2005; Watanabe et al., 2008). In human speech communication, certain disfluencies tend to systematically occur with discourse features and speakers' intentions, which means that the presence of a disfluency, as

well as its type and location, should be used to derive information (Arnold and Tanenhaus, 2011; Shriberg, 1994, 2001).

In all of these studies, disfluencies are treated as part of normal speech, but this is not the only type of disfluency. There are two main ways to look at disfluencies, normally occurring speech disfluencies and disfluencies resulting from communication disorder. In this study, I am only interested in normally occurring disfluencies and I am not investigating clinical disfluencies. Nonetheless, it is important to define clinical disfluencies so as to establish the type of disfluencies that I am not looking at.

2.3 Clinical disfluencies

2.3.1 Communication disorders

A communication disorder can result in a primary disability or can accompany other disabilities. It consists of the impairment in the ability to receive, transmit, and comprehend a message. Communication disorders can be related to speech, language, hearing, or auditory processing. Since the present work focuses on disfluencies, I am only going to go over speech disorders. Speech disorders can be a damage to speech sound articulation, phonology, fluency, and/or voice, and can result in interference of intelligibility. Speech disorders are directly related to clinical disfluencies due to articulation and phonological disorders that can create atypical production of speech sounds due to omissions, substitutions, additions, or distortions. Similarly, voice disorders create disturbances or the absence of vocal quality in terms of pitch, loudness, resonance, and/or duration. Finally, fluency disorders can create interruptions in the flow of speaking by altering rhythm, and creating repetitions of various segments such as sounds, syllables, words, and phrases (American Speech-Language-Hearing Association Ad Hoc Committee on Service Delivery in the Schools, 1993; Anderson and Shames, 2011).

2.3.2 Defining fluent speech

Speech fluency is important to define when looking at clinical disfluencies because it allows us to determine better what disfluent speech is. Normal speech fluency is defined by a number of factors : speech rate, utterance length, rhythm, timing, consistency in the duration of elements, continuity, the quantity of filled and unfilled pauses, and finally, the effort that goes into speech production (Starkweather, 1987). Therefore, disfluent speech contains higher rates of pauses and disruptions to the speech rate and duration consistency ; it also requires more effort to produce speech. The disruptions can consist of various types of clinical disfluencies, further defined in the following subsection 2.3.3.

2.3.3 Various types of clinical disfluencies

Clinical disfluencies, unlike non-clinical disfluencies, occur in the speech of people who have disorders that might be directly linked to speech. For example, individuals who stutter (IWS) have a disorder primarily related to speech, but people who have Down Syndrome or Autistic Disorder Syndrome have a disorder non-primarily related to speech, but which causes them to have speech impediments (Kumin, 1994; Preus, 1972; Sisskin, 2006) .

Clinical disfluencies constitute a broad category of disfluencies. A common distinction among clinical disfluencies is between stutter-like disfluencies (SLD) and non stutter-like disfluencies (NSLD).

Stutter-like disfluencies (SLDs)

Typical SLDs consist of part-word repetitions (sounds or syllables), entire word repetitions, broken words, and finally, audible or inaudible prolongations (Anderson and Shames, 2011).

Stuttering is one of the most common fluency disorders. Developmental stuttering is the most common type of stuttering. It begins in early childhood, affects 5-15% of the population, and persists into adulthood in about 1% of the population (Goberman et al.,

2010). Neurogenic stuttering is a rarer form of stuttering and is generally acquired after trauma to the brain. Psychogenic stuttering is another rare condition where stuttering occurs as a result of an emotionally or psychologically traumatic experience.

There are several diverging definitions of stuttering, and no set universal definition (Goberman et al., 2010). One of the most cited and recognized definition of stuttering is from Wingate's (1964) *A Standard Definition of Stuttering* : "Stuttering is broadly defined as (a) disruption in the fluency of verbal expression, which is (b) characterized by involuntary, audible or silent, repetitions or prolongations in the utterance of short speech elements - namely, sounds, syllables, and words of one syllable. These disruptions (c) usually occur frequently or are marked in character and (d) are not readily controllable. Sometimes the disruptions are (e) accompanied by accessory activities involving the speech apparatus, related or unrelated body structures, or stereotyped speech utterances. These activities give the appearance of being speech-related struggle. Also, there are not infrequently (f) indications or report of the presence of an emotional state ranging from a general condition of "excitement" or "tension" to more specific emotions of a negative nature such as fear, embarrassment, irritation, or the like. (g) The immediate source of stuttering is some incoordination expressed in the peripheral speech mechanism ; the ultimate cause is presently unknown and may be complex or compound" (Wingate, 1964, p. 488).

The cause for stuttering is still unknown, but several studies have found physiological differences between individuals who stutter (IWS) and individuals who do not stutter (IWNS). A few examples of physiological differences are differences in cortical activation (Blomgren et al., 2003), the activity of the basal ganglia (Giraud et al., 2008), or the role of the dopamine system (Giraud et al., 2008; Fetterolf and Marceau, 2013).

In addition to physiological factors, there are also linguistic factors that play a role affecting stuttering. Those linguistic factors consist of several dimensions : prosodic, phonetic, lexical, and syntactic properties. One of the most prominent studies that looked at the linguistic factors of stuttering focused on the specific loci of stuttering moments. The study revealed that stuttering tends to occur more on consonants than on vowels, on sounds in

word-initial position, in connected speech than on isolated words, on content words than on function words, on longer words than on shorter words, on words in earlier position in the sentence than in later position, and finally, stuttering tends to occur more on stressed syllables than unstressed ones (Brown, 1945). A more recent study (Kadri et al., 2011) revealed similar results (illustrated and summarized in Figure 2.1).

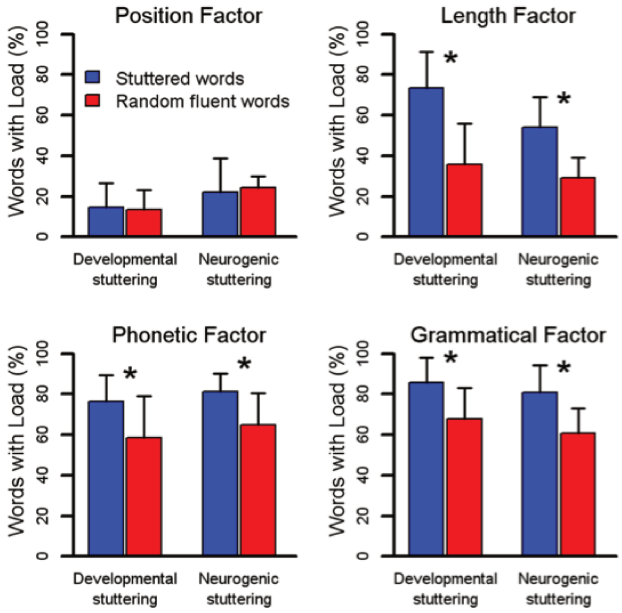

FIGURE 2.1: Linguistic factors in stuttering, after Kadri et al. (2011)

Stuttering is situated on a continuum of disfluencies and is hard to quantify. There is no clear-cut boundary of disfluency rate to determine whether an individual stutters. In addition, stuttering rates vary drastically within and across individuals who stutter depending on several factors such as fatigue, cognitive load, moment of the day, or the use of techniques to avoid stuttering. A study from Noeth et al. (2000) showed that the average IWS have a disfluency rate of about 10% whereas IWNS tend to have a disfluency rate that averages around 2%.

Sluttering is another similar, but less common, SLD speech disorder that starts during

childhood. Cluttering consists of "rapid, dysrhythmic, sporadic, unorganized, and frequently unintelligible speech" (Daly and Burnett, 1996, p.239).

Individuals with Parkinson disease also have SLDs, characterized as Neurogenic, as opposed to developmental disfluencies, because they are caused by a brain disorder (Goberman et al., 2010). Goberman's (2010) study reports stutter like behaviors such as repeated movements and fixed articulatory postures, with within-word and between-word disfluencies.

Individuals with Tourette Syndrome (ITS) also have SLDs such as word repetitions, hesitations, interjections, and prolongations. In addition to speech disfluencies, both ITS and IWS share biological and environmental components such as the genetic factor, a higher prevalence in men, and an increase in symptoms triggered by higher emotional stress (De Nil et al., 2005).

Non stutter-like disfluencies (NSLDs) and disfluencies not directly related to stuttering

Typical NSLDs are multisyllable word repetitions, word-final repetitions, phrase repetitions, abnormal voice modulations and sound swaps (Scaler Scott et al., 2006; Shriberg et al., 2001b).

A study on the clinical disfluencies of three subjects with Asperger Syndrome and Attention Deficit Disorder showed that in addition to SLDs, subjects also had disfluency patterns different from regular IWS. The NSLDs were final-word stuttering, long sporadic stuttering events concomitant with long blocks, phrase repetitions, and revisions (Scaler Scott et al., 2006).

Similarly, individuals with Tourette Syndrome (ITS) have other disfluencies in addition to SLDs. Those disluencies are either non-clinical, filled and unfilled pauses, and interjections ; or clinical NSLDs, such as abnormal voice fluctuations and sound swaps (De Nil et al., 2005).

2.4 Normally occurring disfluencies

Normally occurring disfluencies are present in the speech of individuals who do not have a speech disorder. Normally occurring disfluencies, also referred to as *non-clinical disfluencies*,

are the type of disfluencies that I am interested in. Therefore, from here on I refer to them as *disfluencies*.

Disfluencies have been widely studied in the discursive, cognitive and computational literature, and several classifications have been provided. For instance, Lutz and Mallard (1986) use nine categories to generally classify disfluencies : prolongations, part-word repetitions, word and/or phrase repetition, interjections, revisions, incomplete phrases, dysrhythmic phonations (abnormal laryngeal behaviors including interjected moments of glottal fry and breaks in the production of voicing), incoherent sounds, and fillers. Shriberg (2001) uses a different classification with six types of disfluencies : filled pauses, repetitions, deletions (also called false starts), substitutions, insertions, and articulation errors (also referred to as speech errors).

2.4.1 Why studying disfluencies ?

While they are often treated as noise or irregularities, disfluencies tend to follow specific patterns (Shriberg, 1994). There is evidence that shows that disfluencies can have functions, especially for listeners (Arnold and Tanenhaus, 2011; Fox Tree, 2001; Clark and Fox Tree, 2002; Siu and Ostendorf, 1996).

The purpose of studying disfluencies and how they pattern ranges from improving speech technology performance and naturalness, to understanding human language production, perception, and cognition (Shriberg, 1994). The goals of this dissertation are to understand the role(s) of disfluencies in spontaneous speech, and what discourse and acoustic patterns are associated with them. One of the applications of studying disfluency functions and patterns is to bring information to the development of natural speech systems to improve spontaneous conversation speech understanding (see subsection 2.4.5). Therefore, in this dissertation, I mostly focus on details that relate to these applications and I do not talk exhaustively about models of language production, perception or cognition.

2.4.2 Disfluency regions

Previous authors have identified specific regions within disfluencies. Different parts of the disfluency correspond to different functions and usually have different acoustic properties. In this work, I follow Shriberg's (1994, 2001) model, which identifies three regions and one point : the reparandum (RM), the editing phase, the repair and, the interruption point (IP). The IP is the cutoff point at the end of the reparandum, also considered as the departure point from fluency. The reparandum consists of the entire stretch of speech to be deleted, the editing phase is the region from the IP to the onset of the third region, the repair. The editing phase can be empty or can contain a silent pause, editing phrases or filled pauses (*sorry, I mean, um, uh*). Finally, the repair stands for the resumption of fluency and usually corresponds to the stretch of speech in the RM. Figure 2.2 illustrates the different parts of the disfluency : the reparendum *the cat-*, replaced by the repair *the dog*, with an editing phase from the IP to the repair *uh*.

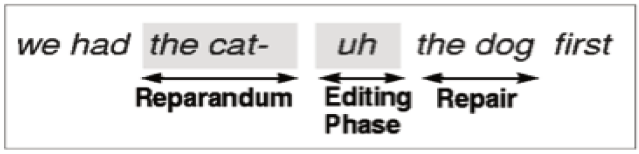

FIGURE 2.2: Disfluency regions, after Shriberg (1994, 2001)

2.4.3 Acoustic properties of disfluencies

Several studies have shown that specific areas of the disfluency are characterized by specific acoustic cues. For instance, disfluencies are commonly marked by the lengthening of rhymes or the lengthening of syllables preceding the Intonation Phrase (IP) (Lickley, 1994; Shriberg, 1994, 2001). Furthermore, filled pauses in English are usually characterized by a long, steady central vowel with little spectral change (Gabrea and O'Shaughnessy, 2000) .

This paragraph summarizes the acoustic properties of disfluencies in spontaneous speech

of American English from Shriberg (2001). Shriberg claims that disfluencies have consequences for the acoustic and phonetic properties of speech. As explained in subsection 2.4.2, disfluencies can be divided into three regions. Each region differs in terms of acoustic properties. Data for her study come from three corpora : Switchboard - a collection of free conversation between humans, AMEX - a collection of human-human air travel dialogs and, ATIS - a collection of human to computer air travel dialog. Most of the phonetic effects in the reparandum are prevalent around the IP. Phonetic effects in this region consist of duration patterns, laryngealization, voice quality and vowel quality. One of the most observed patterns is lengthening of the rhyme or syllable preceding the IP. Shriberg differs lengthening in disfluencies from pre-boundary lengthening. Pre-boundary lengthening is observed in fluent speech as the only cue to disfluency. The motivations to differentiate both phenomena are mostly justified by acoustic differences. Lengthening in disfluencies is much greater than in the lexically fluent sequence and pitch tends to be flat or slowly falling, similarly to filled pauses, mostly when the disfluency is associated with hesitation. The lengthening of the reparandum is usually accompanied by creaky voicing which triggers the 'trailing off' percept with a decrease in amplitude and a drop in pitch. Other phenomena in the reparandum lead to specific acoustic marking. Word cutoffs usually show some laryngealization. The reparandum is also often characterized by alterations in voice quality in words such as *the*, *a* or *to*. *A* and *to* are more likely to be pronounced with their tense vowel forms and *the* is more likely to be pronounced as /ði/ in the reparandum of a disfluency than elsewhere. Unfortunately, Shriberg does not provide numbers to justify her claim in this section. Effects in the editing phase are mainly related to duration. According to Shriberg, unfilled pauses tend to be long and are good cues for disfluency detection. However, the author does not mention a duration number to support this claim. In addition, she mentions that certain unfilled pauses can be shorter than regular pauses. Vowels in filled pauses, however, are much longer than elsewhere. In addition to duration cues, F0 in filled pauses has been shown to be low, gradually falling and related to the surrounding environment. Finally, effects in the repair are usually not prevalent since the repair is the resumption of fluency and therefore most consequences

of the disfluency are observed in the reparandum and editing phase. It has been observed however that when there is a form of contrastive emphasis in the repair, there is usually one or more of the following : an increase in F0, duration or amplitude. This type of acoustic information can be incorporated into automatic speech understanding systems in order to be used as cues for disfluency (see subsection 2.4.5). More background on the acoustic properties of *um* and *uh* is provided in section 3.1 of the current chapter.

2.4.4 Variables that affect disfluency rate

As previously mentionned, about 6% of words are disfluent in conversational American English (Fox Tree, 1995; Shriberg, 1994). Similarly, a study on French broadcast interview archives showed that four types of disfluencies (discourse markers, filled pauses, repetitions and revisions) constitute about 8% of the corpus (Mareüil et al., 2005). Numerous factors affect the rate of disfluencies : cognitive load, utterance length and complexity, task, social context, time constraints, etc. (Moniz et al., 2014; Shriberg, 1994, 2001; Tottie, 2014).

Studies have looked at how gender and task affect disfluency rate. For instance, Lutz and Mallard (1986) conducted a study on 25 female and 25 male undergraduate and graduate students at Southwest Texas State University and found that the rate of disfluencies between males and females differs more in conversation than during reading. In conversation, the disfluency rate of males was 3.2% of words and 2.4% of syllables against 3.0% and 2.7% for females. During reading, the disfluency rate of males and females was the same : 1% of words and 0.8% of syllables. Results of median percentage of total disfluency by category during conversation showed that interjections and revisions were the most common types of disfluencies, and that during conversation the rate of filler-to-word ratio was on average 4 fillers per 100 words. Males and Females used similar amounts of prolongations, part-word repetitions, interjections, revisions and fillers. However, results show that males use more word and/or phrase repetitions (15.2%) and incomplete phrases (3%) than females (10.5% and 0.1%) ; and females use more incoherent sounds (9.5%) than males (0.2%).

A study on the use of disfluencies in European Portuguese in university lectures and

map-task dialogues shows that speaking style affects the production of disfluencies in terms of distributional patterns and prosodic properties (Moniz et al., 2014). Results show that speakers produced more repetitions and fragments in dialogues than in lectures, which the authors attribute to speech planning acts and time constraints corresponding to the two speaking styles. Results also show that disfluencies are shorter in dialogues and that in lectures, speakers use more pitch and energy increases to differentiate disfluency regions and adjacent contexts than in the dialogues.

More background on the variables that affect the rate and the production of *um* and *uh* is provided in section 3.1.

2.4.5 Issues for Speech Technologies

Even though speech technology has significantly improved in recent years, speech processing and understanding still heavily depend on punctuated and fluent input. Spontaneous spoken speech is not fluent, it structurally differs from written speech and, it contains prosodic information (e.g.,the way we say things). Therefore, current systems based on fluent, written or controlled speech usually face challenges when dealing with spontaneous and conversational spoken speech and fail to use part of the information present in the signal.

Spontaneous speech

Spontaneous spoken speech differs from written, read or, planned speech in several ways. Some of the major distinctions that cause issues for speech technology trained or modeled after non-spontaneous speech are differences in sentence structure, the presence of prosodic information, and disfluencies.

Spontaneous spoken speech has sentence-like units instead of typical written sentences. These sentence-like units consist in grammatical, semantically complete, shorter sentences (Liu et al., 2006a). Four subtypes of sentence-like units (statements, questions, backchannels or incomplete sentences) are offered in Strassel (2004). Sentence-like units can consist of only one word or one noun phrase, especially in the case of answers.

Spontaneous spoken speech also contains prosodic information encoded in the speech signal which carries structural, semantic or, pragmatic information. For instance, prosodic cues can disambiguate syntactic structures, indicate questions or statements, speakers' attitudes and, focal information, help detect disfluencies and sentence boundaries (Levow et al., 2014; Liu et al., 2006a; Stouten et al., 2006; Wang et al., 2013).

Another main distinction is that spontaneous spoken speech is not fluent and contains high rates of disfluencies. Disfluencies present a major challenge for spontaneous speech processing and understanding (Shriberg, 2001; Liu et al., 2006a; Stouten et al., 2006; Wang et al., 2013). They disrupt the grammatical flow of sentences and can create word-like elements not necessary in the lexicon of the recognizer or structures that do not map to the Language Model. For instance, the content of edit disfluencies can be edited, repeated or, dropped. For an example of edit disfluency and its different regions, see Figure 2.2. Liu et al. (2006a) propose four subtypes of edit disfluencies : repetitions, revisions, restarts and complex disfluencies, the later consisting of nested or successive disfluencies. Filler words are another group of disfluencies that also disrupt the regular flow of sentences and that are not present in transcripts of written texts or read speech. Filler words consist of filled pauses (FPs), discourse markers (DMs) and, explicit editing terms (Liu et al., 2006a). FPs can be hesitation markers or floor holders, most common FPs in English are *ah*, *eh*, *um* and *uh*. Discourse markers are words that carry discourse and structural information such as *you know, I mean, so, well, like, etc.* Explicit editing terms are disfluencies that mark repairs (see Figure 2.2) and are not an FP or a DM (e.g., *I like <cats->, <sorry>, <dogs>* where *sorrry* is the explicit editing term).

Detecting speech disfluencies

To solve issues caused by disfluencies present in spontaneous spoken speech, several speech understanding systems aim at filtering disfluencies. Removing disfluencies can be necessary or useful to clean up transcripts. The resulting transcripts can be used for downstream natural language processing tasks or to improve transcript readability (Ferguson et al., 2015; Hassan

et al., 2014; Snover et al., 2004). However, recent research shows that certain disfluencies contain discourse or prosodic information that can be used to improve performance of speech processing and understanding systems (Siu and Ostendorf, 1996; Siu et al., 2000; Stolcke and Shriberg, 1996). Therefore, removing disfluencies might not always be the preferred approach. In either case, whether disfluencies are filtered or, preserved and used to extract information, the first step is to detect them.

Different techniques are used for disfluency detection. Some of the most popular techniques are reviewed in this paragraph. Several systems use a mixture of prosodic cues and lexical features. Prosodic cues are used to find the general location of a disfluent event, and focus on identifying the interruption point (IP) (Liu et al., 2006a; Shriberg, 1999). Other systems use models primarily based on lexical features (Snover et al., 2004), which identify the words themselves and the Part-of-Speech tags, without relying on comprehensive prosodic cues. Results show the lexically based algorithm performs comparatively to other algorithms that make heavier uses of prosodic cues. Probabilistic syntactic models are another detection system based on parse structures and a noisy channel model to identify disfluencies (Lease et al., 2006; Zwarts et al., 2010). Other methods include internally and externally informed strategies to handle disfluencies in spontaneous speech (Stouten et al., 2006). The internally informed search strategy consists in letting the Language Model, the lexicon and the acoustic model work jointly to hypothesize the disfluency and to create a context manipulation that decides the path of action for the recognition system ; while the externally informed search strategy mainly consists in an external detector responsible for identifying disfluencies, and associates posterior probabilities to the disfluency segments. Predictive models such as n-gram models are another way to detect disfluencies. N-gram models associate probabilities to sequences of words or entire sentences, where a *n-gram* is a sequence of N words (Honal and Schultz, 2005; Stolcke and Shriberg, 1996; Siu et al., 2000; Zwarts and Johnson, 2011). Discriminative models such as Conditional Random Fields (CRF) are also used to detect disfluencies (Georgila, 2009; Ostendorf and Hahn, 2013; Zayats et al., 2015). CRF models directly model the conditional probability of $p(y|x)$. They efficiently model multivariate out-

puts y while including a large set of input features x for prediction (Sutton and McCallum, 2012). CRF models are a log-linear model for sequential labels, similar to and more powerful than Hidden Markov Models (HMM). Compared to other methods, CRF models yield to better results (Ostendorf and Hahn, 2013). However, state of the art studies show that Neural Network models that incorporate acoustic-prosodic and lexical features show improvements over other methods (Wang et al., 2015; Zhang et al., 2014).

2.5 Synthesis

This chapter provided an extensive review of what disfluencies are, their different types, whether clinical or naturally-occurring, and their acoustic and discourse properties. Clinical disfluencies occur in the speech of people with speech disorders, or other disorders that cause speech impediments, and are commonly divided into two broad categories, stutter-like and non stutter-like disfluencies. Normally occurring disfluencies occur in the speech of individuals who do not have speech disorders, and are the type of disfluencies I am interested in.

Disfluencies are often treated as random speech errors that disrupt the normal flow of speech. Several studies however have showed there are different types of disfluencies, and disfluencies tend to follow specific patterns and can have functions in discourse. The study of disfluencies leads to improvements in speech technology performance and naturalness, as well as our understanding of human language.

Disfluencies span a wide range of topics and can take different forms. In this dissertation I focus on two disfluencies : the markers *um* and *uh*. Even though *um* and *uh* are often grouped under the same label, one of the main goals of this dissertation is to show that they are not the same disfluency, and are therefore studied separately. The following chapter provides an extensive review of the two markers, ranging from their various functions in speech, to their acoustic properties and the challenges they present in speech processing.

Chapter 3

UM AND *UH*

3.1 Introduction

In this dissertation, I focus on two specific disfluencies : *um* and *uh*. The goal of the current chapter is to provide background on the existing research on *um* and *uh*, to point out issues related to these two disfluencies, and to propose ways to address them.

There are three main issues to address. First, whether *um* and *uh* behave in the same way. Second, whether they have systematic functions in spontaneous speech that correspond to specific linguistic environments. The last issue is to determine whether all instances of *um* and *uh* mark discourse information ; or if they are simply the result of speech planning.

3.1.1 Background on um and uh

Um and *uh* are some of the most frequent items in spoken American and British English. A mixed speech corpus analysis shows the combined frequency of *um* and *uh* is greater than 13,000 per million words (pmw) (Biber et al., 1999). Another study on American English, that looks at the first part of the Santa Barbara Corpus, reports 7,500 occurrences of *um* and *uh* pmw (Tottie, 2015a).

Terminology

Multiple terms in the literature are used to designate *um* and *uh* : *hesitations* (Corley and Stewart, 2008), *fillers* (Clark and Fox Tree, 2002; Corley and Stewart, 2008; O'Connell and Kowal, 2005), *filler words* (Arnold and Tanenhaus, 2011), *filled pauses* (Shriberg and Lickley,

1993; Shriberg, 1994, 1999, 2001; Swerts et al., 1998; Watanabe et al., 2008), and *UHM*[1] (Tottie, 2014). According to Clark and Fox Tree (2002), calling *um* and *uh* *filled pauses* assumes that they are simply pauses filled with sounds. However, in order to consider the hypothesis that *um* and *uh* have a discourse function, the authors argue that it is better to call them *fillers*. *Fillers* however are usually used as a more flexible term that refers to *filled pauses* and *discourse markers* such as *I mean*, *you know*, *like*, and *well* (Laserna et al., 2014; Tottie, 2014, 2015a). Tottie (2015a) collapses *um* and *uh* under the same label *UHM* because he considers them to be the same variable. In the present work I follow the terminology from Clark and Fox Tree (2002) and reserve the term *pause* for silent pauses only, and I refer to *um* and *uh* as *markers*. I do not adopt the term *UHM* because I hypothesize that the two markers have different functions in speech.

Um and *uh* have been traditionally treated as disfluencies along the lines of repetitions, false starts, hesitations and pauses. Recent research however has focused on their status, discourse functions and prosodic, acoustic and, syntactic environment.

The rate of *um* and *uh* is affected by several extra-linguistic factors such as cognitive load, task effect (human-human versus human-computer interaction), speaker, listener and planning time allowed to the speaker (Tottie, 2014). Similarly, intra-linguistic factors such as sentence or utterance length and complexity also affect the markers' rate (Shriberg, 1994, 2001; Watanabe et al., 2008).

There is little agreement on the function of *um* and *uh* in spontaneous speech, that is, whether *um* and *uh* are by-products of the planning process or whether speakers use them intentionally. Some of the literature considers *um* and *uh* to be errors or random symptoms in language performance and argues that they should not be considered part of the linguistic signal (Chomsky, 1965). Tottie (2015b) analyzes certain tokens of *um* and *uh* as non-linguistic signal when they function as floor-holder. However, several recent studies support the idea of *um* and *uh* as part of language, and therefore as signal instead of symptom (Arnold and

1. The term *UHM* refers to *um* and *uh* at the same time

Tanenhaus, 2011; Clark and Fox Tree, 2002; Norrick, 2015; Tottie, 2014, 2015a,b).

Furthermore, authors disagree on the status of *um* and *uh*, that is, whether they should be considered disfluencies like hesitations or filled pauses, or if they should be considered English words belonging to the category of interjections (Clark and Fox Tree, 2002; Norrick, 2015), or pragmatic markers (Tottie, 2014). At this stage, it is still unclear what all the functions of *um* and *uh* are, and whether the two markers are interchangeable. This study does not address the status of *um* and *uh* (i.e., word *vs.* disfluency *vs.* pragmatic marker). Instead, it investigates the environment in which *um* and *uh* occur and looks for patterns to see whether they are systematically associated with certain phenomena and if they behave acoustically differently.

3.1.2 Study goals

The purpose of this study is to shed light on the functions of *um* and *uh* in order to know which markers bring information to the discourse, and whether *um* and *uh* behave differently from each other. This information can then be used to establish which markers should be filtered or used by natural language processing systems (see subsection 2.4.5). In this study, I investigate the effects of speaker and listener gender, position of the marker, speakers' attitudes towards what they are saying and, the acoustic, prosodic and, syntactic environments of the markers on the distribution and properties of *um* and *uh* in spoken spontaneous speech.

3.1.3 Functions of um and uh

Several studies show that filled pauses are used in various ways and can be attributed specific functions in discourse. Items such as *um* and *uh* are documented to have functions that range from filled pauses to floor holders, discourse markers or even interjections. Several patterns such as features of discourse, speakers' intentions and prosodic environment regarding the use of *um* and *uh* indicate that they are not random and that they are very likely to take part in the linguistic signal. These systematic occurrences are one of the main

arguments in favor of *um* and *uh* being part of the linguistic signal. The following subsections review studies on their various functions.

How listeners use um and uh

Several studies show that listeners can derive information from the use of *um* and *uh*. For instance, some studies investigate how speaker disfluency affects on-line accessibility of referents during reference comprehension and show that listeners' expectations regarding the speakers' reference can be influenced by disfluencies : listeners prefer to interpret old information when the speaker's instructions are fluent whereas they prefer interpreting new information when the instruction are disfluent (Arnold et al., 2004, 2007; Arnold and Tanenhaus, 2011).

Watanabe et al. (2008) look at whether filled pauses affect listener's predictions about the complexity of upcoming phrases in Japanese. The study focuses on native Japanese speakers and native Chinese listeners to see whether they get similar cues from filled pauses. Findings show that filled pauses are cues to complex phrases and that effects of filled pauses on non-native listeners depend on their fluency. Results are coherent with findings from Clark and Fox Tree (2002) (see subsection 3.1.3).

Based on the assumption that *um* and *uh* tend to occur in predictable environments, such as before unfamiliar words, before words with low token frequency, or before newly mentionned words, Kidd et al. (2011) show that children pay attention to *um* and *uh* and that they use the information carried by the markers during online spoken word recognition. These markers are used to anticipate upcoming referents. Their results also suggest that this ability to use disfluencies seems to arise around the age of two.

In sum, studies show that both children and adults use *um* and *uh* as cues for processing difficulty and online accessibility of referents, and therefore indicate that the markers have a function in discourse and that they are part of the linguistic signal.

How speakers use um and uh

Two studies that show speakers seem to have control over their use of *um* and *uh* in different registers suggest that the makers should be considered as signal rather than just speech errors. The studies report that on average, speakers use *um* and *uh* less in formal registers than in informal register (Duez, 1982; Schachter et al., 1991). In Schachter et al. (1991) speakers use *um* and *uh* less in formal registers despite the fact that they produce more pauses, speak more slowly, and in theory have more planning time than in the informal setting. A more recent study, however, presents different results : speakers use *um* and *uh* more in non-private environments (e.g. offices and classrooms) than in private settings like homes. The study also reveals that planning time and cognitive load seem to be important factors in the frequency of *um* and *uh* since even in private settings, complicated discussion topics lead to higher rates of markers (Tottie, 2014).

Further evidence for *um* and *uh* as signal is their use as floor holders to indicate that speakers are still engaged in their speech act and that their turn is not over (Kjellmer, 2003; Shriberg, 2001; Tottie, 2015a). Findings from a study of turn-taking in the Santa Barbara Corpus of Spoken American English (SBCSAE) reveal that *um* and *uh* can have both a turn-holding and turn-yielding function (Tottie, 2015a).

Clark and Fox Tree (2002) argue that speakers have control over the use of *um* and *uh*, that speakers can plan them and, that the markers have the status of English words. The authors define word as *"linguistic units that have conventional phonological shapes and meanings and are governed by the rules of syntax and prosody"* (Clark and Fox Tree, 2002, 75). The markers are also categorized and analyzed as English interjections (Clark and Fox Tree, 2002; Norrick, 2015). Interjections are *"not integrated into the grammatical structure of the utterances/clauses ; they can stand alone as independent units in their own right ; they tend to have functions rather than meaning ; they differ by national, regional and personal variety"* (Norrick, 2015, 249). Norrick (2015) categorizes *um* and *uh* as phatic interjections due to their phatic function illustrated in example (1). ""

(1) Madonna : oh, well you can move it in there, put it on the floor. I brought some candy.
Earl : *um*.
Doug : for after dinner.
Madonna : no need for me to start on things, I'll get in a hurry and everything will be read [before]
Earl : [before the] turkey
(LSWE-AC 144801) from (Norrick, 2015)

Clark and Fox Tree (2002) support that similarly to interjections, *um* and *uh* have a basic meaning and implicatures. The authors define the core meaning of the markers as interjections : *uh* indicates a minor upcoming delay and *um* a major one. Other uses derive from this basic meaning and therefore are implicatures of announcing minor or major delays. Findings from their study on the London–Lund corpus of British English (LL) (Svartvik, 1990) show that lengthened *um* and *uh* are more likely to co-occur with pauses. However, findings from Tottie (2015b) also contradict part of Clark and Fox Tree (2002)'s study and show that in American English 330/957 occurrences of *um* and *uh* did not occur with pauses and about half of the tokens did not introduce delays but instead ended them. Similarly, findings from O'Connell and Kowal (2005) in a corpus study of TV and radio interviews between Senator Hillary Clinton and six interviewers show that most occurrences of *um* and *uh* are not followed by silent pauses. O'Connell and Kowal conclude that *um* and *uh* fail to reliably indicate minor or major delays. In addition, they reject the proposal that *um* and *uh* are interjections because they cannot constitute a turn by themselves and because they occur in different positions relative to pauses. Their findings indicate that the markers tend to be preceded by a pause in initial position, whereas interjections tend to occur between pauses. Norrick (2015) argues that *um* and *uh* can constitute turns and stand by themselves (as illustrated in (1)) and that interjections, like *um* and *uh*, also occur in initial position. Tottie (2015b) does not support the claim that *um* and *uh* are interjections in oral speech but he recognizes

their interjection-like status in restricted and rare written cases to convey ironic euphemism as in (2) or polite disagreement as in (3).

(2) Obama is more, *um*, seasoned
Barack Obama's... closely shorn hair appears to be increasingly gray.
Washington Post, August 28, 2008, from (Tottie, 2015b).

(3) ... Senator Richard Shelby of Alabama... "The market will view these firms as... implicitly backed by government." *Um*, senator, the market already views those firms as having implicit government backing, because they do... (Paul Krugman, Op-Ed column, New York Times) from (Tottie, 2015b).

In fact, Tottie argues that a vast proportion of bare *um* and *uh* (without preceding or following pauses) should be considered symptom instead of signal. Even though uses of the markers might have originated as hesitation markers, the author argues that certain uses of *um* and *uh* have evolved as real words functioning as pragmatic markers such as *well*, *I mean*, *you know*, and *like* (Tottie, 2014). Tottie proposes the following analysis of the two markers :

> UHM operates on a gradient, originating in spoken language as a symptom of processing difficulty or need for planning time, and proceeds via uses as a pragmatic marker to occasional specialized and deliberate use as a signal with quasi-word status (Tottie, 2015b, 51).

3.1.4 Acoustics of um and uh

Results from Shriberg (2001) show that *um* and *uh* have distinct acoustic properties. The vowel of the markers is much longer than the same vowel elsewhere and the pitch of *um* and *uh* tends to fall systematically between the previous peak and the bottom of the speaker's pitch range.

Intonation units (IU) are one of the most basic units of spoken speech. They are charac-

terized by a single intonational contour (Nespor and Vogel, 1986) and range from clauses to single words. Clark and Fox Tree (2002) identify three key positions in IUs : (I) the boundary, (II) after the first word excluding *um* and *uh* and, (III) later. Findings from their research show that there are more markers in location I than II, and in location II than III, consistent with the idea that planning should have the most effect in position I.

These specific acoustic features could be a sign of salience for the markers and could be used in language models to differentiate *uh* from the article *a* for instance. Other implications for speech technology include improving disfluency detection and separating disfluencies from regular words in duration modeling to avoid skewing word duration.

3.1.5 Um and uh in language processing

As mentioned in subsection 2.4.5, many speech understanding systems filter disfluencies with the goal of improving system performance. However, Stolcke and Shriberg (1996) show that removing disfluencies increases the perplexity of surrounding words. In addition, several studies show that filled pauses like *um* and *uh* contain discourse and prosodic information such as marking linguistic units and restart boundaries or indicating hesitation when a speaker holds the floor (Siu and Ostendorf, 1996; Siu et al., 2000). *Um* and *uh* can also predict neighboring words and tend to precede words with a lower token frequency (Stolcke and Shriberg, 1996). The context of the filled pause can also be used to predict whether the filled pause is worth skipping, depending on its position (e.g., sentence initial vs. medial) (Siu et al., 2000). Siu and Ostendorf (1996) show that the position of the marker affects the perplexity and that depending on the marker and the position, different treatments lead to different results. For instance, skipping sentence-medial *uh* reduces perplexity but not necessarily for other markers or other positions.

Such studies suggest that we need to look at the environment, and the acoustic, prosodic, and discourse properties of the markers to determine whether a disfluency carries relevant information.

3.1.6 Um vs. uh

In most of the literature, *um* and *uh* are collapsed under the same category, or label. For instance, Tottie refers to the two markers as one and the same, that she calls *UHM*, and describes them as two variants of the same variable (Tottie, 2015b).

However, several studies find differences between the two markers. One of the main differences is the position of the marker. Clark and Fox Tree (2002) found that *um* is more likely to occur at utterance boundary and *uh* is more likely to occur utterance medially. They also found that *um* indicates major upcoming delays, whereas *uh* indicates minor ones. Even though the latter results have been contested by other others, it is still interesting to note differences between the two markers to further investigate whether they have the same functions.

Based on the principle that *um* and *uh* are different from other disfluencies because they have pragmatic and listerner-oriented functions, two recent studies investigate the use of the two markers by individuals with Autism Spectrum Disorders (ASD) and show differences in use of the markers for children with ASD. Individuals with ASD typically display pragmatic language impairments. Irvine et al. (2015) compare the production of *um* and *uh* in spontaneous speech between three youth groups : individuals with ASD who have pragmatic language impairments, individuals with ASD whose language disorders are resolved, and individuals with typical development (TD). Results show the rates of *uh* did not differ between groups, but participants with ASD who have pragmatic language impairments produced significantly fewer *ums* than the other two groups. The authors suggest that the production of *um* correlates with autism symptom severity and that the marker has as a pragmatic, listener-oriented function. These results are consistent with another study on the production of *um* and *uh* by children age 4-8 with ASD during the autism diagnostic observation schedule. Gorman et al. (2016) compare the production of *um* and *uh* by three groups of children, children with ASD, with TD and with specific language impairment (SLI). Individuals with SLI typically have structural issues with language whereas individuals with ASD have both

structural and pragmatic impairments. Similarly to Irvine et al. (2015), results show that children with ASD use significantly fewer *ums* than children with TD.

These studies suggest that *um* and *uh* do not have the same function and therefore are not the same marker. In this study I separate the two markers. One of the goals of this dissertation is to identify how different they are and if all *ums* are different from *uhs* or if there is some overlap.

3.2 Synthesis

To summarize, there are different views regarding the status and functions of *um* and *uh* in the literature, ranging from symptoms of language performance to markers of hesitation, floor holders, interjections and pragmatic markers. I posit that these uses are not mutually exclusive and that the classification and function of the markers highly depends on the immediate prosodic and discourse context, the level of involvement of the speaker and the acoustic features of the marker. The aim of this study is therefore to explore the discourse environments and the properties of *um* and *uh*, in order to see if we can find information from systematic or abnormal occurrences.

O'Connell and Kowal (2005) argue that there are empirical issues with Clark and Fox Tree (2002) because the LL corpus was annotated by coders instead of automatic transcriptions, which means that pause duration reflects the annotators' perception of the pause length rather than its actual duration. Similarly, Tottie (2015b) uses the Santa Barbara Corpus of Spoken American English (SBC) with perception-based marking for pause transcription. O'Connell and Kowal (2005) address this issue by acoustically measuring the duration of pauses in Praat (Boersma and Weenink, 2015) for their study. We know that uses and rates of *um* and *uh* are highly idiosyncratic, extra-linguistic and intra-linguistic dependent (Shriberg, 2001; Tottie, 2014). Therefore, I anticipate some major issues with the corpus used in O'Connell and Kowal (2005). It contains a limited number of speakers, six interviewers and one politician, and 70% of *um* and *uh* (600 out of 861) come from the same speaker, Hillary Clinton. In addition, journalists and politicians are trained speakers and therefore

are not representative of everyday speech.

3.3 Issues to address

There are three main questions to address regarding the markers *um* and *uh*. The first is to test whether *um* and *uh* are the same variable. That is, whether they function in the same way and whether they occur in the same environments. To address this question, I study the two markers separately. The second question is to find out whether they have systematic functions in spontaneous speech that correspond to specific linguistic environments. To address this, in various experiments I investigate the linguistic environments of the markers, as well as their acoustic properties depending on a wide range of variables, in two corpora. The last issue is to determine whether all instances of *um* and *uh* are meaningful to the discourse (i.e., planning errors or filled pauses vs. floor holders or pragmatic markers).

To answer these questions, I conduct a wide range of analyses and experiments. I first look at how *um* and *uh* vary depending on several discourse and linguistic variables (Chapter 6). I also look at variables that affect transcription errors of *um* and *uh*, to find out if listeners perceive more *um* than *uh*, and how differently from other words (Chapter 7). Then I look at how various properties of *um* and *uh* correlate with speaker attitude (*stance*, see Chapter 4), to find out whether their presence, position, and acoustic properties can predict the stance of the speakers, if there are internal structures of *um* and *uh*, and if acoustic and discourse features relevant to the two markers improve automatic stance classification (respectively Chapters 8, 9, and 10).

The attitude or the sentiment of the speakers is referred to as *stance* in this study. Stance is commonly referred to as *sentiment*, *evaluation*, or *emotion*, and is different from the study of *sentiment analysis*. Chapter 4 provides an exhaustive review on the topic of *stance*, defines the concept used in the frame of this study, and reviews relevant literature to this work.

Chapter 4

STANCE

4.1 Introduction

Stance is an overt communicative act that uses language. It is expressed by a stancetaker, concerning the object of the stance act (Du Bois, 2007). Broadly speaking, stance refers to "the speakers' subjective attitudes toward something" (Haddington, 2004, p. 101) and involves sociocultural, dialogical, linguistic, and intersubjective dimensions (Du Bois, 2007).

In section 4.2 I first review the various definitions of stance, and explain which definition I apply to this study. In section 4.3 I explain why I am interested in stance and I review the state of the art on automatic stance recognition on spoken data.

4.2 Definitions

Stance is a broad concept that can be defined in several ways depending on the researchers' interests. It is studied in several fields such as Anthropology, Education or Sociology, as well as in various subfields of Linguistics such as Corpus Linguistics, Cognitive Linguistics or Sociolinguistics. The core idea of stance is that it focuses on the function of language, and on the context in which it is used, to represent the social and pragmatic functions of language (Englebretson, 2007). The term *stance* is therefore used in various ways, and different researchers using this term do not necessarily refer to the same thing. Conversely, several other terms are also used to refer to stance, such as *subjectivity*, *evaluation*, or *sentiment*. (Englebretson, 2007, p. 16) summarizes *subjectivity* as "broad self-expression" and *evaluation* as "subjectivity with a focus". Biber's definition is one of the many definitions of *stance*, and summarizes it concisely as "personal feelings, attitudes, value judgments, or assessments" (Biber et al., 1999, p. 966). According to Du Bois, "the act of taking a stance necessarily

invokes an evaluation at one level or another, whether by assertion or inference" (Du Bois, 2007, p. 141). Key components of stance are evaluation, positioning and alignment, including "sociocognitive relations of objective, subjective and intersubjective intentionality" (Du Bois, 2007, p. 162).

The expression of stance can be qualified as *subjective* when looking at the function of language for individuals, or as *intersubjective* when looking at its function between individuals (Scheibman, 2007). Englebretson argues that *stance* is a conceptual entity which can be researched, that people actively engage in *stance*, and stance-taking takes place in discourse. Stance can be expressed by grammatical and lexical patterns such as certain combination of verb types and tenses, or by use of adverbials, evaluative adjectives, modals, and expressions (Hunston and Thompson, 2000; Scheibman, 2002). Stance can also be expressed by gestures or acoustically. Studies show that different levels of stance reliably affect pronunciation aspects linked to prosody such as vowel duration, speech rate or intensity (Freeman et al., 2014a; Freeman, 2014). Another study on the prosody of *yeah* shows that pitch and intensity increase as stance strength increases. The study also shows that stance polarity affects the acoustic signal, as the pitch and the intensity of negative *yeahs* is slightly higher than for positive or neutral *yeahs* (Freeman et al., 2015b).

In this dissertation, what I consider by *stance* is best defined by "the speakers' subjective attitudes toward something" (Haddington, 2004, p. 101). I consider two dimensions of stance : its strength (no stance, weak stance, moderate stance, or strong stance), as well as its polarity (neutral, positive, or negative), as annotated in the ATAROS corpus (see section 5.3.4 in Chapter 5) (Freeman, 2015).

4.3 Why stance ?

In automatic recognition, stance is often referred to as *sentiment*, *emotion*, *opinion*, *subjectivity*, or *private state*, a term that denotes beliefs, judgments, and evaluations, with functions such as *experiencers*, *attitudes*, and *targets* (Quirk, 1985; Wilson and Wiebe, 2005). The notion of *stance* is slightly different from *sentiment analysis*. Sentiment analysis, also

referred to as *opinion mining* or *subjectivity analysis* (Pang and Lee, 2008), is used to refer to the automatic analysis of the evaluation of a topic. Traditionally, the concept of *sentiment analysis* has been used to refer to polarity (negative or positive), but is now used in broader terms to refer to more general concepts such as opinion and subjectivity. The concept of stance, as defined in this study, differs slightly from sentiment analysis. It is measured in two dimensions, strength and polarity, and from an enunciative point of view takes the position of the enunciator and the co-enunciator into account.

Most of the research on automatic stance recognition, also called *automatic subjectivity recognition*, is conducted on text-based speech, using primarily lexical and syntactic features (Pang et al., 2002; Somasundaran and Wiebe, 2009). A few studies have focused on automatic recognition of stance in spoken speech, mainly based on n-grams, lexical, and structural features (Godfrey et al., 1992; Murray and Carenini, 2009). Murray and Carenini (2009) focused on detecting subjective sentences in spontaneous speech in multiparty speech, in order to label them for polarity. Their results indicate that the use of n-grams with shifting levels of lexical instantiation improved the performance over prior methods. As previously mentioned, recent studies (Freeman, 2014; Freeman et al., 2015b) show that stance in spontaneous spoken speech is marked by several elements not present in text, and (Somasundaran et al., 2006) found that inter-annotator agreement on opinion annotation is higher when transcribers have access to both spoken and written data than when they have access to transcripts only. A study from (Levow et al., 2014) on automatic recognition of stance in spontaneous speech takes into account lexical, speaking styles and prosodic features, to train a classifier to label stance behavior in a boosting framework. Their results show that lexical information alone (i.e., word unigram features) lead to the best accuracies (71-80%), while speaking style and prosodic features lead to lower performance of the system. Prosodic and speaking style features combined lead to 55.2% accuracy for stance strength, and the features alone lead to 71% accuracy, above most common class assignment. These results also show that prosodic and speaking style features improve the accuracy of stance polarity classification more than stance strength classification. A followup experiment, however, shows that the manually an-

notated punctuation masks the effect of prosody. That is, when punctuation features are omitted for stance strength classification, adding prosodic and speaking style features to the lexical features increases the system's accuracy from 61.5 to 63%. The study also reports on the first 25 word unigrams selected by the classifier across all folds for the recognition of the degree of stance. One of the most interesting findings from (Levow et al., 2014) in the light of this dissertation is that *um* figures among the 25 word unigrams. This suggest that *um* is an important feature of stance strength marking, and therefore further corroborates the fact that *um* has functions in discourse, including stance strength marking.

Based on the findings reported in the current chapter and on Chapter 2, I will investigate the role and the acoustic properties of the markers *um* and *uh* in stance marking in the ATAROS corpus (see section 5.3.4 for more information on the stance annotation in the corpus). But, first, Chapter 5 introduces the corpora used for our data.

Chapter 5

CORPORA

5.1 Introduction

The two corpora used in this study are the ATAROS corpus and the Switchboard corpus. Unlike the ATAROS corpus, which is rather small, the Switchboard corpus contains a lot more data, necessary for statistical analysis or training data for machine learning. In addition, Switchboard represents a larger sample of the population, with speakers from 8 dialect areas although unbalanced (see section 5.4.2). The ATAROS corpus focuses on speakers from the Pacific Northwest exclusively, who work on the same collaborative tasks, and was designed to elicit various degrees of involvement, with high quality recordings.

5.2 Corpus choice rationale

My goals are to investigate the discourse environment and the prosodic properties of the markers *um* and *uh* in spontaneous speech, as well as their relationship to syntactic structure, and whether they are involved in stance marking. For more information on stance marking see section 4.1 in Chapter 4, and for more information on the ATAROS corpus annotations on stance see section 5.3.4 in the current Chapter.

I therefore need corpora that have spontaneous speech, the presence of *um* and *uh*, high-quality recordings, the markings of stance, part of speech tagging, parses and, prosodic annotations. No single corpus contains all of these elements together, therefore I use a combination of corpora. The use of several corpora raises questions on the topic of interoperability, especially on how to compare findings across datasets.

The ATAROS corpus has overt markings of stance, tasks that elicit various degrees of speaker involvement and, disfluencies (including *um* and *uh*). I therefore use it to look at the

functions of *um* and *uh* relative to stance, as well as their distribution depending on various discourse variables such as degree of involvement. Similarly, since part of the Switchboard corpus has syntactic, prosodic, and disfluency annotations, I use it when looking at the relationship of disfluencies to prosodic and syntactic environments in which *um* and *uh* happen. In addition, Switchboard has different transcript versions, with different levels of transcription precision, which allow comparing how speakers noticed *um* and *uh*, and how the transcription errors vary depending on various factors.

The two corpora used for this study consist of recorded spontaneous speech with high quality recordings of different speech tasks. A combination of the two corpora is used to answer different questions depending on the type of speech and the type of annotation available for the corpus.

5.3 The ATAROS corpus

The ATAROS (Automatic Tagging and Recognition of Stance) corpus is an audio corpus of collaborative tasks between dyads (groups of two speakers) with high quality recordings (see section 5.3.2 for more information on the recording conditions). The corpus was designed to elicit different degrees of involvement and various degrees of stance (see section 4.1 in Chapter 4), and to look at the acoustic signal of stance-taking.

5.3.1 Corpus design

The collaborative tasks between dyads consist of unscripted conversations. Each dyad executes five tasks designed to trigger different levels of involvement and changes in stance. Tasks are divided into two groups. Each group contains a set of about fifty tokens that aim at representing the main vowels of Western American English. The first group contains the Map, Inventory and Survival Tasks and the second group consists of the Category and Budget Tasks. In each group, the first task is designed to elicit a baseline with stance-neutral conversation (Map and Category Tasks) ; speakers discuss how the fifty items are arranged in a different order. The Inventory, Survival and Budget Tasks are collaborative

decision-making tasks and were respectively designed to elicit increasing levels of involvement and stance-taking. The Inventory Task consists of arranging items from a superstore into a new inventory and is designed to elicit the lowest level of involvement and weak stances. This task usually triggers polite proposals and suggestions between the two speakers. In the Survival Task, speakers have to discuss which items to leave or keep in a survival scenario. It is designed to elicit a higher degree of involvement and higher levels of stances than the Inventory Task. In the Budget Task, subjects have to imagine they are on a county budget committee and have to decide which items should be cut. This task is designed to elicit the highest degree of involvement and stronger stances. In this task, speakers demonstrate more sophisticated discussions and negotiations and are likely to refer to personal experience to corroborate stances (Freeman et al., 2014b). This study only looks at tokens from the Inventory and Budget Tasks because they respectively elicit the lowest and highest degrees of involvement and stance (Freeman et al., 2014b).

5.3.2 Recording conditions

Recordings were made in a sound-attenuated booth at the University of Washington in Seattle with head-mounted AKG C520 condenser microphones. An XLR cable connected them to a separate channel in an M-Audio Profire 610 mixer outside the booth. Recordings were saved as 16-bit stereo WAV-file at a 44.1 kHz sampling rate. The major advantages of this corpus are the high-quality recordings and the use of head-mounted microphones which allows controlling for acoustic intensity and improves signal-to-noise ratio.

5.3.3 Speakers

The corpus currently contains 34 recordings of same and mixed gender dyads[1]. At the time of the study, 17 analyzable recordings were annotated at the coarse level for the Budget and Inventory Tasks. Dyads consist of strangers matched approximately by age. Speakers

1. At the time of publication (Freeman et al., 2014b) there were only 26 annotated dyads.

are Native English speakers from the Pacific Northwest and range from age 18 to 75. The total duration of recordings per dyad for the five tasks ranges from 40 to 80 minutes, with an average of 60 minutes. In this study, I am looking at a subset of the corpus : 17 dyads, 9 mixed gender dyads, 5 female-female and 3 male-male dyads (19 females and 15 males). Speakers' age ranges from 20 to 70 with a mean of 35 for women and 39 for men.

5.3.4 Transcriptions and stance-related annotations

All annotations are made in Praat (Boersma and Weenink, 2015) by trained annotators. Annotations for each speaker in the Inventory and Budget Tasks consist of five levels, each respectively annotated in separate tiers : phones, words, utterances, coarse stance annotation and, fine stance annotation. The utterance tier is manually transcribed and words are transcribed orthographically according to conventional American spelling. The word and phone tiers are created from the utterance tier with forced alignment using the Penn Phonetics Lab Forced Aligner (P2FA) (Yuan and Liberman, 2008). Two types of pauses are annotated in the corpus : pauses under 500ms are marked within an utterance by 2 periods '..' and pauses over 500ms are marked as silence 'sp'. *Um* and *uh* are transcribed as 'UM' and 'UH' in the word tier, the former showing nasality. In the utterance tier the two tokens are transcribed as 'um' and 'uh'. When they are appended to other words, they are transcribed attached to the word separated by a short dash without space (e.g., *uh-oh*).

The coarse annotation tier indicates four levels of stance strength : no stance (marked 0, (see example (1)), weak (marked 1, see examples (2) and (3)), moderate (marked 2, see examples (4) and (5)) and strong stance (marked 3, see examples (6) and (7)). The tier is also annotated for three levels of stance polarity : positive (marked '+', see (5)), negative (marked '-', see (3) and (6)) and neutral (not marked see (1),(2),(4) and (7)). Tokens for which annotators could not tell are marked with *'x'*. Zero stance corresponds to factual statements or questions, backchannels, reading or conversation managers such as *'Okay'* or *'next'*. Weak stance corresponds to superficial agreements, opinion solicitations, solution offers, mild encouragements or opinions. Moderate stance is essentially a stronger version

of weak stance with questioning of other's opinion, strong personal credibility and excited exclamations. Strong stance consists of emphatic and stronger manifestations of weak and moderate stance and can be loaded and/or emotional. Positive polarity is marked on items expressing agreement, confirmation, encouragement or intonation that conveys positivity. Negative polarity is marked on items expressing disagreement, questioning of other's opinion and intonation that conveys negativity. Neutral polarity is associated to items marked with 0 stance. To summarize, there are two stance dimensions, strength and polarity. For instance, example (3) illustrates negative weak stance, and example (5) illustrates positive moderate stance.

(1) Stance 0 : And then socks. QUAL breathy VOC laugh Um.
(NWF089-NWM053-3I, 77.946sec)

(2) Stance 1 : As a, um, personal hygiene item.
(NWF106-NWF107-3I, 50.330sec)

(3) Stance 1- : Uh, yeah. QUAL reluctant VOC breath
(NWM061-NWM060-6B, 460.596sec)

(4) Stance 2 : Um. Oh, but this is sweets, too. QUAL muttering Oh, that's *baking though. Important distinction.
(NWF089-NWM053-3I, 42.257sec)

(5) Stance 2+ : I - uh, well.. I'd say tools, yeah.
(NWM055-NWF093-3I, 794.019sec)

(6) Stance 3- : Oh but *eggs! Well, uh, heck yes. Gosh. That was an obvious one that I missed. VOC laugh
(NWF090-NWF091-3I, 622.403sec)

(7) Stance 3 : Um, for *this arrangement we have.. identified no area for clothing.
(NWF106-NWF107-3I, 118.891sec)

There is a finer-grained annotation of stance available in ATAROS but for statistical power reason I do not use it. For more information on this annotation refer to (Freeman et al., 2015a).

For stance strength and polarity, each task is annotated by an annotator and then reviewed or corrected by another. Uncertainties noted by the first annotator are reviewed and if the second annotator remains uncertain, a thrid annotator intervenes as a tie breaker. Results of the weighted Cohen's kappas with equidistant penalties on this annotation method show high inter-rater agreement : 0.87% for stance strength labels and 0.93% for stance polarity labels (p=0) (Freeman, 2015).

5.3.5 Spurts

Since this study focuses on both the discursive context and the prosodic characteristics of *um* and *uh*, I investigate token frequency of the markers in relation to discourse structure based units. Utterances are closely related to the discourse structure from an intonation point of view. However, since intonation units are not transcribed in the ATAROS corpus, the position of *um* and *uh* is analyzed relative to the spurts. In this study, spurts follow the same definition as in ATAROS. They correspond to utterances between pauses greater than 500ms, and are used as a unit for stance annotations, although stance acts can sometimes span several spurts, or spurts can contain several stance annotations (Freeman, 2015).

5.3.6 Measurements

Polarity and strength of stance are collected from the coarse tier of each TextGrid. Acoustic data are automatically collected by a Praat script that runs through the 17 sound files and TextGrids. The script collects acoustic measurements at the midpoint of the vowel : pitch, intensity and, F1 and F2 ; as well as word and vowel duration, based on boundaries and transcriptions from the phone and word tiers. Settings for the script are 0.01sec time step, 5 formants, 5500Hz maximum formant, 0.025sec window length and a minimum pitch

of 60Hz. I used threshold of 30ms minimum for vowel duration since it is difficult to get accurate measurements for any token with a vowel shorter than that.

5.3.7 Advantages of the ATAROS corpus

This corpus has several advantages. First, since it was designed to analyze the acoustics of stance-taking, the corpus contains recordings of tasks eliciting different levels of speaker involvement and, stance strength and polarity are transcribed by trained annotators (see subsection 5.3.4). Another advantage of the subset of this corpus used in the present study is that it contains an adequate number of non-professionally trained speakers (17 speakers), which is more representative of how people speak in conversational English, unlike recordings of journalists or politicians. In addition, the recordings have high acoustic quality. Pauses are acoustically measured instead of being transcribed based on duration perception. Speakers are recorded with a head-mounted microphones, which means that intensity level is consistent within speaker.

5.4 The Switchboard corpus

Switchboard is a corpus of spontaneous conversational speech. It is a multi-speaker database of telephone bandwidth speech, designed to train and test speech algorithms with time aligned word transcriptions, and a variety of annotations depending on the version or the subset of the corpus (Godfrey et al., 1992). Some of the most commonly used annotations include Treebank3, Mississippi State, the disfluency annotations, and Switchboard NXT. Switchboard is a broadly used corpus, especially in a wide range of studies that look at speech disfluencies such as Ostendorf and Hahn (2013); Shriberg (1996, 2001); Stouten et al. (2006); Zayats et al. (2014), etc.

5.4.1 Corpus design

The Switchboard corpus consists of telephone conversations between paid volunteers of both sexes from a variety of backgrounds and representing the main American English dia-

lects. The goal of the corpus is to elicit natural, spontaneous speech. Annotators rated the naturalness of the corpus at 1.48 on average, on a scale 1-5 ; 1 representing "very natural" and 5 "forced or artificial sounding" (Godfrey et al., 1992). The average naturalness score suggests that the protocol aim was met.

The Switchboard-1 Release 2 version contains 2,400 conversations, collected automatically, without human intervention. Participants first engage individually with a computer before being in communication. Once in communication, participants are introduced a conversation topic and given a chance to introduce each other before starting the recording. Each subject is recorded synchronously during the conversation in a separate channel. The speech signal is collected directly from the telephone network with a system that allows high quality recordings (Godfrey et al., 1992).

Conversation durations range between 1.5 to 10 minutes, with an average of 6.5 minutes (Calhoun et al., 2010). Godfrey et al. (1992) round the amount of material to around 250 hours of conversation, and about 3 million words. All conversations of the corpus have time-aligned detailed transcriptions.

5.4.2 Participants

A total of 543 speakers, 302 males (56%) and 241 females (44%), participated to at least one conversation. Demographic data was collected for all participants and stored in an Oracle database. Participant information consists of caller ID, sex, birth year, dialect area and education level. Participant age varies from 20 to 60, with 26% of speakers between 20 and 29, 34% between 30 and 39, 21% between 40 and 49, 16% between 50 and 59, and finally 2% between 60 and 69. Participant's dialect is based on where the speaker grew up for the first 10 years of their life, and categorized into 8 dialect areas : south midland (29%), western (16%), north midland (15%), northern (14%), southern (11%), NYC (6%), mixed (5%), and New England (4%). The reason why there are more speakers from the south midland dialect is due to the fact that many participants are TI employees, people connected to the employees, or local people. Education level is categorized in 5 levels : less than high school (2.6%), less

than college (7.2%), college (57%), more than college (32.5%), and unknown (0.7%).

In addition to participant data, information about the call is also automatically entered into the Oracle database : date and time of the call, length of the conversation, area code and telephone number and, any other relevant information (Godfrey et al., 1992).

5.4.3 Transcripts

There are several transcripts of the corpus. The original transcript, the slightly modified Treebank3 transcript (Marcus et al., 1993) and the later corrected MS-State transcript (Deshmukh et al., 1998). Several annotations of the Switchboard corpus were made by different research groups and in different formats, linked to different word transcripts of the conversations.

Treebank3

The Treebank3 transcripts consits of 1126 annotated conversations from the original Switchboard release. The transcripts contain segmentations of each turn into utterances, part of speech tags on each word and, annotated disfluencies. Disfluency annotations identify the reparandum, the interruption point and the repair of the disfluency, according to Shriberg's (1994) annotations.

Um and *uh* are marked as *filler* {F ...}, under the category of non-sentence elements, as illustrated in (8).

(8) B : {Actually, } [I, + {F uh, } I] gues I am <laughter>. / {F um, } it just seems kind of funny that this is a topic of disccussion. / {F uh, } I do, {F uh, } some, {F uh, } woodworking myself <noise>. {F uh, } in fact, I'm in the middle of a project right now making a bed for my son. /
A : {F uh, } {F uh } I see stuff in craft galleries for five hundred dollars /
From Meteer and Taylor (1995)

The utterance units are sentence-like chunks, called *slash units*. Even though most speaking turns do not necessarily consist of complete sentences, the slash units are considered to be complete utterances and are parsed at S (i.e., highest level : sentence). In other words, they might not be sentential, but they are still interpreted as complete spoken utterances (illustrated in example (9)). For incomplete turns, where the speaker stopped midstream, slash units are marked as incomplete (see example (10)). Fillers can also make up a turn of its own, even when interrupting another turn (see example (11)). For further details on slah units and disfluency annotations, refer to Meteer and Taylor (1995).

The Penn Treebank3 syntactic release is a subset of 650 conversations which contain full syntactic parses. For further details on the syntactic annotations see Marcus et al. (1993).

(9) A : Yeah, / right. /
B : Yeah. / Kind of jack of all trades, master of none. /
A : I went there, / we have Home Depot out here. /
From Meteer and Taylor (1995)

(10) A : ... I don't thing they're always necessary. / If you put enough patience into, -/
B : – with it / {C and, } {F um, } -/
From Meteer and Taylor (1995)

(11) A : he's pretty good. / He stays out of the street. / {C and, } {F uh, } if I catch him I call him / {C and } he comes back. / {D so, } [he, + he's] pretty good about taking to commands [and + –
B : {F um } /
A : – abd] things. /
From Meteer and Taylor (1995)

Mississippi State

The Mississippi State transcripts, called MS-state transcripts, are a clean up project that hand checked and corrected the transcripts from the 1126 Treebank3 conversations. The MS-state transcripts are more accurate than the Treebank3 transcripts, and they contain word alignments with the audio file for each word beginning and end. The word alignments were made automatically with partial manual correction.

5.5 Summary

To summarize, the two corpora present different advantages. As previously mentioned, I use a combination of ATAROS and Switchboard to answer different questions in this dissertation, depending on what advantages they present. When looking at stance, I exclusively look at ATAROS since it is annotated for stance, while for investigating transcription errors I exclusively use Switchboard because it has different transcription versions. I mention in each chapter which corpus is used, and whether a combination of both is used. I also use various subsets of the two corpora depending on the research question of the chapter, and each subset contains different counts of *um* and *uh*. The subset of the corpus, the total number of markers, and the motivations for using the corpus or the subset are all specified in each chapter.

In addition to presenting different advantages, cross-corpora studies shed light on the need for more universal transcriptions and annotations that would allow more reliable comparisons, and that would increase the robustness of findings. Furthermore, cross-corpora studies are important because they inform us on the metastructure of discourse across a variety of activities that help us better understand the mechanisms of spontaneous speech.

Chapter 6

DISTRIBUTION OF *UM* AND *UH* IN ATAROS AND SWITCHBOARD

6.1 Introduction

One of the main goals of this dissertation is to find out whether *um* and *uh* are separate entities. To answer this question, I first explore in this chapter how the two markers vary depending on various discourse variables to determine if they are used interchangeably and in systematic ways. I explore their distribution in two corpora : the ATAROS corpus and the Switchboard corpus. The goal of using two corpora is to see if the distribution of *um* and *uh* extends across different speech activities. In ATAROS, pairs of strangers from the Pacific Northwest collaborate on tasks that were designed to elicit different degrees of speaker involvement, while Switchboard consists of telephone conversations on a given topic between strangers where dialect is not constrained. Spontaneous speech consists of a number of categories and aspects that vary depending on the speaker's activity, such as daily conversations, arguments, collaborations, expression of opinions and feelings, story telling, etc. The two corpora used in this study are not intended to be representative of all types of spontaneous speech, but rather to represent two variants of spontaneous speech.

The two corpora differ in terms of activity, format, and transcription conventions. I am therefore not able to compare all variables across the two corpora. The variables present in ATAROS and Switchboard are speaker and dyad gender. Speaker gender corresponds to the gender of a given speaker while dyad gender corresponds to the gender of two speakers in a conversation, i.e. speaker and listener. In addition I also look at three corpus specific variables. The variable specific to ATAROS is speaker involvement, determined by the speech task, and the variables specific to Switchboard are the naturalness of the conversation and speaker

participation (i.e., the number of conversations speakers participated in). The dependent variables are the rates and the duration of *um* and *uh*.

Based on findings from the literature and on qualitative observations of the two data sets, I predict the following hypotheses : H1) *um* and *uh* are separate entities and their distribution differs depending on discourse variables ; and H2) the distribution trends of the two markers are similar across the two corpora.

The second section of this chapter summarizes a few key features of the two corpora used in this analysis. The analysis is divided into two sections, one for each corpus. Section 6.3 looks at how the rates and the duration of *um* and *uh* vary depending on speaker gender, task, and dyad gender in the ATAROS corpus. Section 6.4 investigates the effects of speaker and dyad gender, naturalness, and speaker participation in the Switchboard corpus, and section 6.5 summarizes the results and discusses the trends across the two data sets.

6.2 Methodology

6.2.1 ATAROS

ATAROS is a corpus designed to look at the acoustic signal of stance-taking with high-quality audio recordings and unscripted conversations between dyads (groups of two speakers). The subset used here contains 64,352 words. Certain conversations were discarded due to transcription or recording issues.

One of the interesting features of ATAROS is that it allows for investigation of speaker involvement. Speaker involvement is determined by the task the speakers are performing during the recording. Each task in the corpus is designed to elicit a certain degree of involvement from the speakers. In this study I use the Inventory Task (IT) and the Budget Task (BT) because they respectively elicit the lowest and highest degrees of involvement (Freeman et al., 2014b). For more information on the corpus design, the speakers, the transcribers and the measurements, see section 5.3 in Chapter 5.

6.2.2 Switchboard

The version of the Switchboard corpus used in this experiment is the Mississippi State transcripts, a hand checked and corrected version of the Treebank3 transcripts. This data set consists of 1,443,003 words across 1126 conversations between 384 speakers who participated in 1 to 25 conversations.

One of the advantages of the Switchboard corpus is that transcribers assigned a naturalness score on a scale 1-5 to indicate whether the transcriber finds the conversation natural between the two speakers (Godfrey and Holliman, 1993). A low naturalness score indicates that a conversation sounds natural and a high score means that it sounds artificial. The average naturalness score over the 1126 transcriptions used in this experiment is 1.44 (against 1.48 reported in (Godfrey et al., 1992)), which means that overall the conversations are rated as very natural. For more information on the corpus design, the speakers, the transcribers and the measurements, see section 5.4 in Chapter 5.

6.3 The distribution of um and uh in ATAROS

The data for this experiment consists of 64,352 words, 29,920 in the Inventory Task (IT) and 34,432 in the Budget Task (BT). The total speaking time for the Inventory Task is 8,292 s., with 4,550 s. for women and 3,742.9 s. for men. The total speaking time for the Budget Task is 9,580 s. with 5,482.5 s. for women and 4,097.8 s. for men (see Table 6.1). The speaking time duration is computed by excluding intervals marked as *silence* (sp) which represent pauses greater than 500ms. Speaking time duration is greater in the Budget Task than in the Inventory Task, which is not surprising since the Budget Task has more words than the Inventory Task. The average Inventory Task duration is 243 s. with a range of 67-448 s. and the average Budget Task duration is 281 s. with a range of 55-578 s. Both the average duration and the duration range are greater in the Budget Task than in the Inventory Task. The average speaking time per speaker across task and gender is 262 s. but varies from 92-513 s.. The average speaking time for men is 249 s. in the IT and 273 s. in

the BT, and the average speaking time for women is 239 s. in the IT and 288 s. in the BT.

In the Inventory Task women spoke 15,718 words and men 14,202 words, and in the Budget Task women spoke 19,194 words and men 15,238 words. Table 6.1 summarizes the total speaking time and the total amount of words for each gender in each task. However, 19 women vs. 15 men participated in the recordings. Therefore, I look at the average number of word across speakers for each gender as opposed to the total number of words to determine if a gender speaks more than the other. Figure 6.1 plots the effect of task and gender on the average number of words (left) and on the average speaking time (right) across speakers. The right-side figures shows that on average, women speak longer than men in both tasks, and to a greater extent in the BT, although gender does not have a significant effect ($p > 0.05$). The left-side figure, however, shows that even though women speak longer than men in both tasks, men speak more words than women ($p > 0.05$), and to a greater extent in the BT. However, the significance test shows that word duration is significantly longer for for women (287 ms.) than for men (266 ms.) ($p < 0.001$), which could explain why on average women speak longer than men but with less words. These results show the importance of looking at the effect of gender and task on speaking time and number of words spoken, and also show the importance to look at the two measures because they take different aspects into account.

TABLE 6.1: Total speaking time and total number of words for each task (IT and BT) and each gender (women and men) in the ATAROS corpus

task	gender	total time (*s.*)	total words
IT	W	4,550	15,718
IT	M	3,742.9	14,202
BT	W	5,482.5	19,194
BT	M	4,097.8	15,238

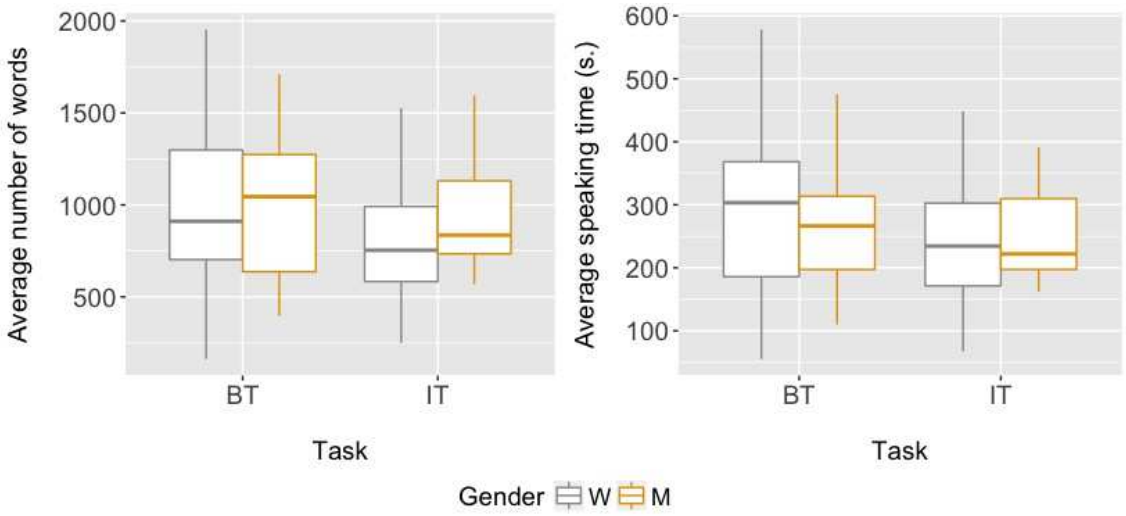

FIGURE 6.1: Average number of words spoken (left) and average speaking time in s. (right) across speakers for each gender and each task

Correspondingly, Freeman (2015) reports that the task length is less variable for the Inventory Task. Freeman's results show that speaking rates, computed in vowels per second

(vps), are significantly faster in the Budget task than in the Inventory Task, and that men have a slightly higher rate than women in both tasks. Additional results show that speaking time varies more when speakers are more involved (i.e., in the BT) and varies more for women than for men.

Speaking time and total number of words show that men and women speak more in the Budget Task than in the Inventory Task. Therefore, since the Budget Task elicits a higher degree of speaker involvement, these measures suggest that speaking time increases as speaker involvement increases.

6.3.1 Task and Gender effect on the rate of um and uh in ATAROS

The corpus used in this study contains 1,065 tokens of the markers *um* and *uh*, which represents 1.65% of words in the corpus (64,352 words in total), with 595 *ums* (0.92% of words in the corpus) and 470 *uhs* (0.73% of words in the corpus). The total number of words is calculated from the word tier of each speaker for each task (17 TextGrids and 34 word tiers, one per speaker) by adding the total number of intervals not marked as silence (pauses greater than 500ms). The proportions of *um* and *uh* are calculated within gender and task in order to look at the effect of task and gender on the distribution of the two markers. The proportions are computed by adding the total number of occurrences of a given token (*um* or *uh*) divided by the total number of words spoken in each task by each gender to normalize the counts. Table 6.2 summarizes the number of tokens in each task for each gender. Results show that in both tasks women use more *ums* (0.88%) than *uhs* (0.54%) whereas men have similar proportions of *ums* and *uhs* (0.97% vs. 0.96%). Results also show that men generally use more *ums* and *uhs* (1.93%) than women (1.42%) across task. According to the frequency of *um* and *uh* in the two different tasks, proportions show that as speaker involvement increases (from the Inventory to the Budget Task), women use .08% more *ums* and 0.21% less *uhs* whereas men use 0.39% more *ums* and 0.37% more *uhs*. When the two categories of markers are collapsed, results show that men tend to use more markers (1.54 to 2.30%) as involvement increases whereas women tend to use fewer markers (1.49 to 1.36%).

TABLE 6.2: Frequency of *um* and *uh* for each gender and each task in the ATAROS corpus

gender	task	*um*	*uh*	total words
W	IT	132 (0.84%)	102 (0.65%)	15,718
W	BT	176 (0.92%)	85 (0.44%)	19,194
W	Total	308 (0.88%)	187 ((0.54%)	34,912
M	IT	110 (0.77%)	109 (0.77%)	14,202
M	BT	177 (1.16%)	174 (1.14%)	15,238
M	Total	287 (0.97%)	283 (0.96%)	29,440

These results indicate substantial differences in the frequency distribution of *um* and *uh* for women, but not men. In addition, the results suggest that we should look at the markers individually, since degree of involvement (task) shows a different effect between *um* and *uh* for women with an inverse tendency, but not for men. These results corroborate findings from Freeman (2015) where men use 1/3 more disfluencies in the Budget Task than in the Inventory Task, based on measurements of *um* and *uh*, truncated words and repetitions between speakers per speech span.

6.3.2 Marker duration in ATAROS

Word duration is significantly longer for women than for men ($p < 0.001$). The average word across the corpus is 287 ms. for women against 266 ms. for men. The average duration of any word is 277 ms. in the Inventory Task and 278 ms. in the Budget Task, with no significant difference reported by the t-test ($p > 0.05$). The duration of any word is measured across all words that are not *um* or *uh*. Duration analyses of *um* and *uh* show that both markers are shorter in the Budget Task, with an average of 484ms for *um* and 335ms for *uh*, against 539ms for *um* and 347ms for *uh* in the Inventory Task, illustrated in Table 6.3. The average duration of *um* varies more across task than for *uh* and the significance test for task effect

over token duration reports a significant effect for *um* ($p < 0.01$) but not for *uh* ($p > 0.05$). Duration results also show that *ums* are longer than *uhs*, which is expected since *um* has two phones against one for *uh*. Figure 6.2 plots the duration of *um* and *uh* for each task and gender. The effect of task is greater for men than for women, and is greater for *um* than for *uh*. The significance test for the effect of task on word duration within gender shows that for men, involvement significantly affects token duration for *um* ($p < 0.01$) but not for *uh* ($p > 0.05$). Hence, for men *ums* are significantly shorter in the Budget Task than in the Inventory Task. The test reports no significant difference in token duration between tasks for women ($p > 0.05$). Furthermore, *um* is longer for men (520 ms.) than for women (494 ms.), while *uh* is longer for women (343 ms.) than for men (339 ms.), but the difference is not significant for either marker.

Duration results are consistent with results from Freeman (2015), that show speaking rates are significantly faster in the Budget task than in the Inventory Task. *Um* and *uh* are both shorter in the Budget task than in the Inventory task, and task and gender have a greater effect on *um* than *uh*.

TABLE 6.3: Average word duration in *ms* for each task in the ATAROS corpus

task	*um*	*uh*	any word
IT	539	347	277
BT	484	335	278

6.3.3 Dyad Effect in ATAROS

This subsection investigates the effect of dyad gender (i.e. the gender of a pair of speakers) on the rate and the duration of *um* and *uh*. Because raw counts of *um* and *uh* across speakers show a wide range of variability (see columns "*um* range" and "*uh* range" in Tables 6.4 and 6.5), the raw counts are normalized by computing the average of the sum of *um* and *uh*

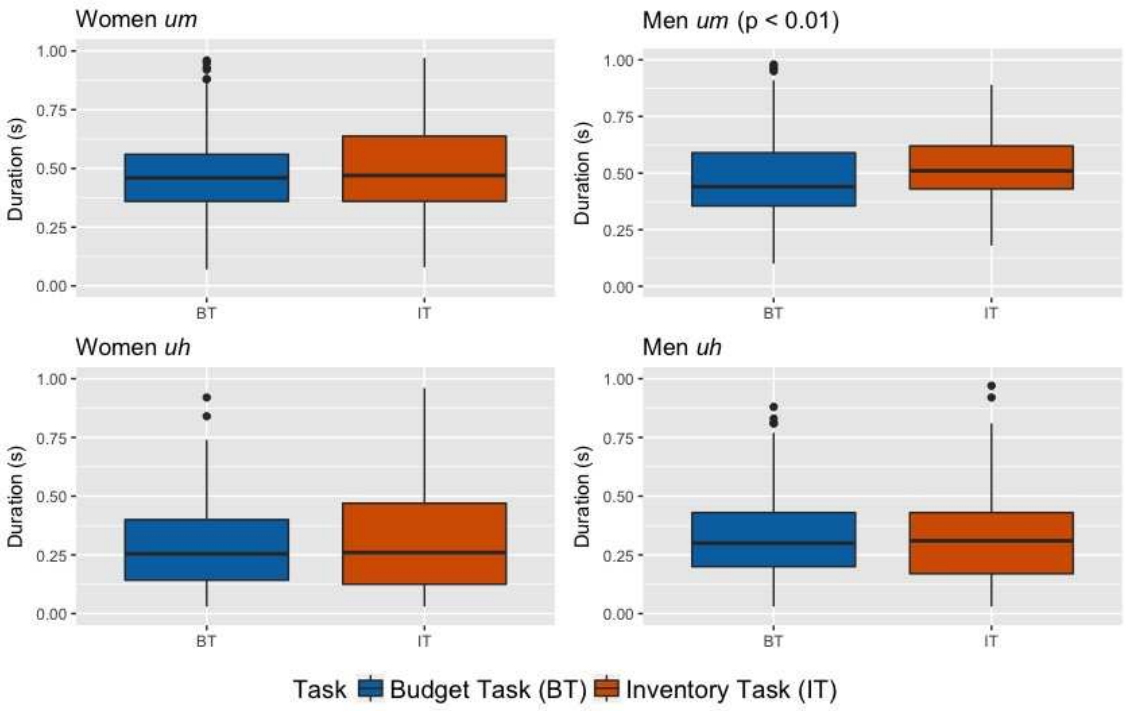

FIGURE 6.2: Task effect on duration of *um* and *uh* for each gender in the ATAROS corpus

for speakers, divided by the number of words they speak. The rates allow comparing the production of *um* and *uh* compared to how much a speaker actually speaks. Both tables summarize the number of dyads for each gender combination as well as the mean, range, and average rate for each marker. Note that the mixed-gender dyad is represented twice, once for each direction ; men talking to women (M-W) and women talking to men (W-M). Table 6.4 shows the rate of *um* is 2 times bigger and more variable in mixed-gender dyads than in same-gender dyads (see column *um* rate). Table 6.5 shows that speakers use more *uhs* in men-men dyads (M-M), as indicated by the rate column, and that there is more variability in the use of *uh* than in the other dyads (see *uh range* column). However, it is important to note that there are only three M-M days, and that more data is needed to confirm this trend, especially due to high inter-speaker variability in the use of markers.

TABLE 6.4: Average number, range, and rate of *um* for each dyad gender in the ATAROS corpus

dyad gender	dyad quantity	*um* mean	*um* range	*um* rate
M-W	9	24	1-87	0.01
W-M	9	23	5-72	0.012
M-M	3	11	1-22	0.006
W-W	5	10	4-35	0.006

TABLE 6.5: Average number, range, and rate of *uh* for each dyad gender in the ATAROS corpus

dyad gender	dyad quantity	*uh* mean	*uh* range	*uh* rate
M-W	9	13	1-31	0.006
W-M	9	10	3-24	0.007
M-M	3	27	14-56	0.014
W-W	5	9.5	2-23	0.005

Figure 6.3 plots the effect of dyad gender on the rates (left) and on the duration (right) of *um* and *uh*. The plots show the average across speakers within each dyad, and the error bars plot the average plus and minus the standard deviation. The left-side plot echoes Tables 6.4 and 6.5, and shows the high rate of variability across speakers. The right-side plot summarizes the average duration of the two markers across speakers grouped by dyad, and shows that there is more variation in the duration of *um* than in the duration of *uh*. The longest *ums* (0.663 s.) are in men-men dayds while the shortest ones (0.475 s.) are used in men-women dyads, and the longest *uhs* (0.375 s.) are used in women-men dyads while the shortest ones (0.312 s.) are used in women-women dyads. Similarly to the left-side plot, the error bars show

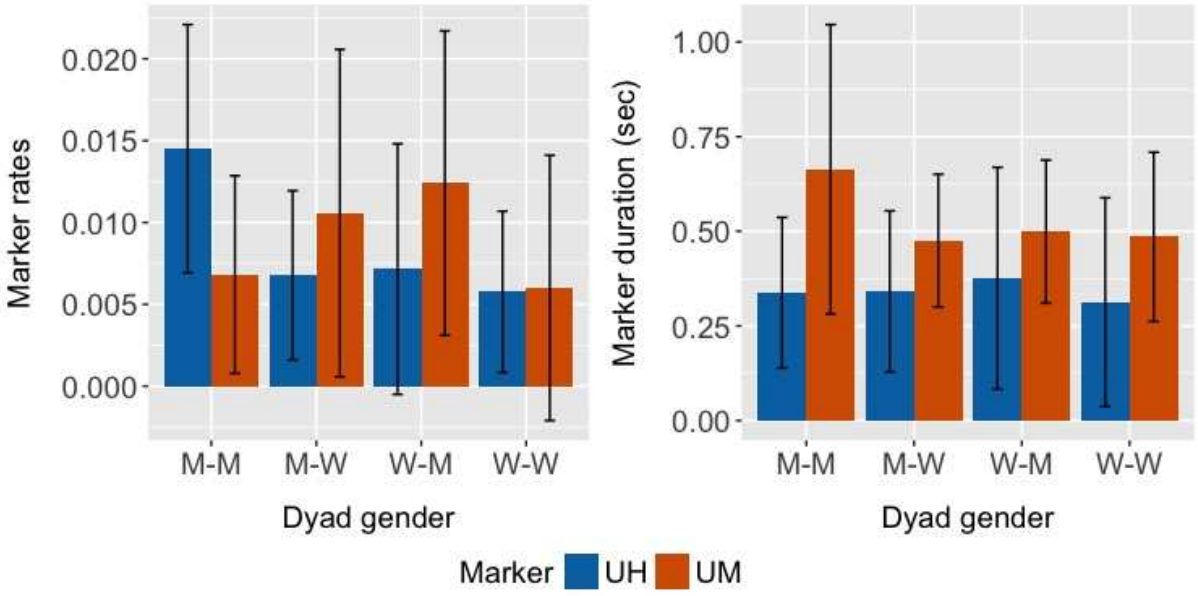

FIGURE 6.3: Effect of dyad gender on the rates (left) and on the duration (right) in *s.* of *um* and *uh* in the ATAROS corpus

the variability in the duration of the two markers across speakers for each dyad. There is a lot of overlap across the dyad categories for both marker rates and duration. This means that dyad effect is highly variable, most likely resulting from the high variability in the speakers' production of *um* and *uh*, and the speaking rate at which they speak depending on the task or the gender.

One of the main conclusions from this analysis is that there is a high variability across speakers in terms of use of *um* and *uh*. Even though there is a lot of variability in the data, the results suggest that speakers use more *ums* in mixed-gender dyads than in same-gender dyads. In the case of *uh*, the results indicate that the highest and lowest rates of markers are found in same-gender dyads. These findings require further investigation due to the small number of same-gender dyads, especially men-men days. Furthermore, there is no systematic gender dyad effect on the duration of the two markers. Consistent with other results from

this study, findings from this experiment on the ATAROS corpus show that *um* and *uh* are affected in different ways. Note that these results differ from results on the rate of disfluencies in Shriberg (2001) that show men use more disfluencies than women, and speakers use more disfluencies with men listeners than with women listeners. These discrepencies may result from the fact that the rate of *um* and *uh* in dyads differ depending on the speech activity. Furthermore, these results could indicate that filled pauses do not behave like all disfluencies, and that looking at them separately can lead to different results.

6.4 The distribution of um and uh in Switchboard

I first look at the general distribution and duration of words in the corpus in order to get a baseline to compare how *um* and *uh* behave relative to other words in the corpus, before looking at the effect of gender, naturalness and speaker participation on the rates and the duration of *um* and *uh*.

Conversation length is measured in two ways. One way is by looking at the number of words spoken either in a conversation, within gender, or within speaker. The other way is to measure the duration by adding the speaking time of speakers (i.e., ignoring pauses) within conversation, gender, or speaker.

6.4.1 Conversation length and duration of all words in Switchboard

The data for this part of the experiment consist of a total of 1,454,919 words across 1,125 conversations with 26,821 unique words. The whole data set contains 1,456,006 words and 1126 conversations but 1 conversation is excluded (1087 words) due to lacking information about a speaker. The average number of words per conversation is 1,293.3 with a minimum of 220 and a maximum of 2,655 words. The average number of words per speaker is 646.6, with a minimum of 34 and a maximum of 1,845 words. Men spoke a total of 644,300 words, with an average of 626.8 words per conversation, a minimum of 87 and a maximum of 1,701. Women spoke 810,619 words in total, 663.4 on average, 34 minimum and 1,845 maximum. These numbers, summarized in Table 6.6, show that there is a large range of variation in

the number of words spoken per conversation. The range of words spoken per conversation is 2,435 words, with a range greater for women (1,811) than for men (1,614) by about 10%. These numbers indicate that there is more variability in terms of spoken words by women than by men, and that women spoke more than men, respectively 810,619 vs. 644,300 words.

TABLE 6.6: Number of words across the Switchboard corpus, per conversation, speaker, and gender

	conversation	speaker	men	women
average	1,293.3	646.6	626.8	663.4
min	220	34	87	34
max	2,655	1,845	1,701	1,845

Table 6.7 summarizes the speaking time per conversation, speaker and gender. Speaking time is computed by adding the duration of words spoken in a conversation, by speaker, or by gender, therefore excluding pauses and silences. The total duration of speaking time across the corpus is 371,098 s. (103.1 hours or 6,185 min). The average speaking duration is 329.9 s. (5.5 min) per conversation and 164.9 s. (2.7 min) per speaker. The total speaking time for men is 160,475.1 s. (44.6 hours or 2,674.6 min), with 156.1 s. (2.6 min) on average, 22.8 s. minimum and 406.7 s. (6.8 min) for maximum speaking time. The total speaking time for women is 210,622.9 s. (58.5 hours or 3,510.4 min), with 172.4 s. (2.9 min) on average, 12.4 s. minimum and 425.8 s. (7.1 min) for maximum speaking time. The average speaking time and the speaking time range show that women and men speak similar amounts of time per conversation, 2.9 min vs. 2.6 min respectively, and with a similar amount of variability.

The average duration of words in the corpus is 255 ms on average across all conversations. Words are significantly shorter for men than for women (249 ms vs 259 ms) as shown by

TABLE 6.7: Speaking time in *s.* across the Switchboard corpus, per conversation, speaker, and gender

	conversation	speaker	men	women
total	371,098	371,098	160,475.1	210,622.9
average	329.9	164.9	156.1	172.4
min	66	12.4	22.8	12.4
max	630.6	425.8	406.7	425.8

the t-test results ($p < 0.001$). The longest words across the corpus are 4.221 s. and 4.073 s. because of laughter accompanying the words *yeah* and *government*. I purposefully exclude minimums and maximums for duration since the shortest words are truncated words, and the longest ones have laughter or other para-linguistic features.

TABLE 6.8: Average duration in ms of words per conversation and gender

	conversation	men	women
average	255	249	259

6.4.2 Um and uh across conversations, speakers and gender in Switchboard

In total, there are 10,784 *ums* in the corpus, and 30,187 *uhs*. *Um* therefore represents 0.74% of all words and *uh* 2.07%. There are therefore 2.8 times more *uhs* than *ums* in this data set, which means that speakers use *uh* almost three times more than *um* in this corpus.

The marker *um* is present in 1,076 conversations and *uh* is found is 1,123 conversations. There are on average 10 *ums* per conversation with a maximum of 53, while there are on average 26.8 *uhs* per conversation with a maximum of 103 *uhs*, which is expected since there are 2.9 more *uhs* than *ums*. Similarly, speakers use on average 3 times more *uhs* than *ums*. What is more interesting, however, is that women use more *ums* than men (0.84% vs. 0.62% respectively) while men use more *uhs* than women (2.76% vs. 1.53% respectively), see Table 6.9. The percentages represent the proportions of *ums* and *uhs* spoken compared to the total number of words (810,619 words in total for women, and 644,300 words in total for men).

Table 6.10 shows the average duration of the two markers within conversation, speaker and gender. Results show that on average, the two markers are significantly shorter for women than men ($p < 0.001$). These results are interesting since on average word duration is longer for women than for men, which means that *um* and *uh* behave differently than the average word.

TABLE 6.9: Number of *ums* and *uhs* across the Switchboard corpus, per conversation, speaker, and gender

	conversation	speaker	men	women
total *um*	10,784	10,784	4,010 (0.62%)	6,774 (0.84%)
average *um*	10	5.8	5.2	6.3
max *um*	53	53	53	41
total *uh*	30,187	30,187	17,761 (2.76%)	12,426 (1.53%)
average *uh*	26.8	13.9	17.6	10.7
max *uh*	103	92	79	92

TABLE 6.10: Duration in s. of *um* and *uh* across the Switchboard corpus, per conversation, speaker, and gender

	conversation	speaker	men	women
average duration of *um*	0.428	0.428	0.435	0.423
average duration of *uh*	0.304	0.300	0.302	0.299

6.4.3 Dyad gender in Switchboard

Figure 6.4 plots the average rates (left) and the average duration (right) for each dyad of the two markers *um* and *uh* depending on dyad gender. The average rate results show that *um* and *uh* are affected differently by dyad gender, and unlike for other variables, the rates of *uh* vary more than for *um* with regards to dyad gender. The lowest and highest rates of *uh* are for same-gender dyads, and there is very little difference for mixed-gender dyads. The lowest rates are between women speakers and listeners (W-W) and the highest rates are between men speakers and listeners (M-M). The rates of *um* are inverse from *uh*. The largest difference is between same-gender dyads, the lowest rates are between men-men dyads (M-M) and the highest rates are between women-women dyads (W-W). The right side plot shows the average duration of each marker depending on dyad gender and shows little to no variation for either marker. The error bars show the mean plus and minus the standard variation for each dyad, and show a homogeneous variability for each marker and each dyad. These results indicate that dyad gender does not affect the duration of *um* and *uh* but that it affects the production of *uh* more than *um*, with an inverse effect for the two markers.

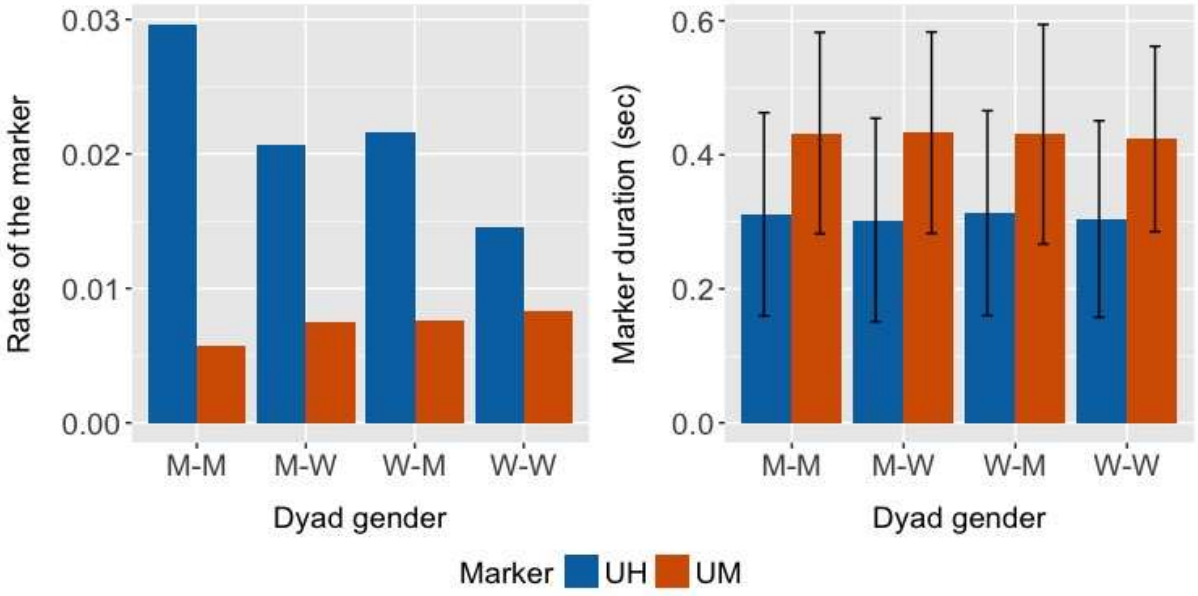

FIGURE 6.4: Effect of dyad gender on the rates (left) and on the duration (right) in *s.* of *um* and *uh* in the Switchboard corpus

6.4.4 Naturalness of the conversation in Switchboard

Since naturalness is a rating of the conversation, I only look at the effect of naturalness at the conversation level. Note that only 1 conversation has the rating 5, which means that the conversation sounds artificial. Table 6.11 summarizes the number of conversations for each naturalness rating, and shows that more than half of the conversations are rated as natural.

TABLE 6.11: Number of conversations for each naturalness rating (1 - natural ; 5 - artificial) in the Switchboard corpus

naturalness	1	2	3	4	5
number of conversations	754	260	96	14	1

Figure 6.5 illustrates the effects of naturalness ratings on the average number of words spoken per conversation (left) and on the average speaking time per conversation (right). In both cases, conversations rated with level 2 of naturalness (i.e., rather natural) correlate with the highest averages of words spoken and speaking time. The overall trend shows that longer conversations, both in terms of words spoken and speaking duration, are rated more natural than shorter ones, with a peak for the second level of naturalness. It is important to remember however that there are only 14 conversations with a naturalness rating of 4, and 1 with a rating of 5 (see 6.11). It is not surprising naturalness has a similar effect on the average number of words and the average speaking time per conversation since the two measures are two ways of measuring conversation length.

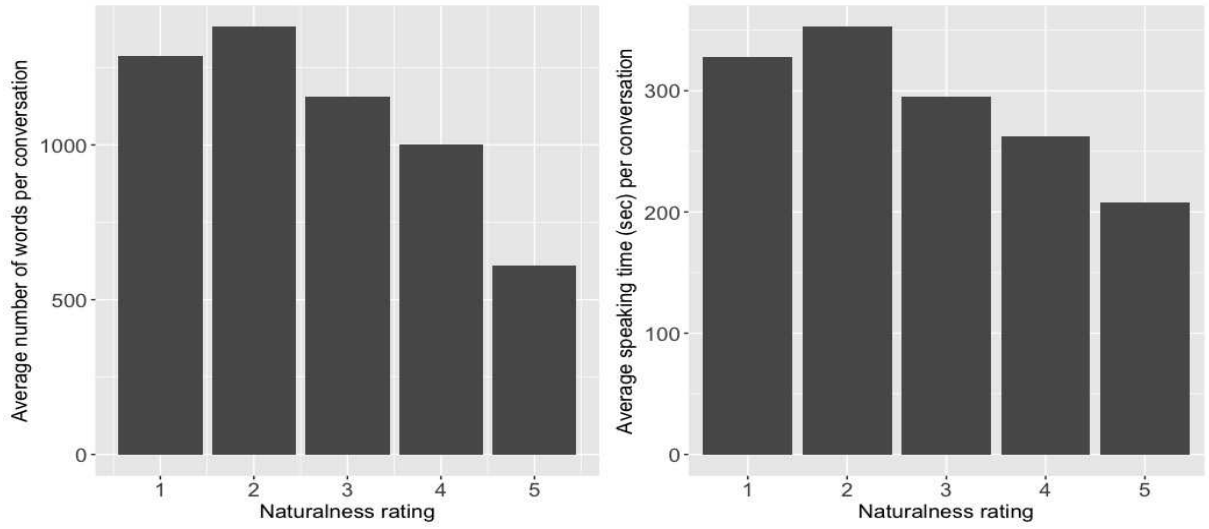

FIGURE 6.5: Effect of naturalness ratings on the average number of words and the average speaking time in *s.* per conversation in the Switchboard corpus

Figure 6.6 plots the effects of naturalness ratings on the average duration (left) and the average rates (right) of *um* and *uh* per conversation. Even though ratings 4 and 5 are plotted, since only 1 conversation has a rating of 5 and only 14 have a rating of 4, I only consider ratings 1, 2, and 3 for this section. Results for ratings between 1 and 3 show the average

marker's duration and rate have similar trends. Naturalness rating between 1 and 3 do not affect the marker's duration. Similarly to other words, conversations with naturalness rating 2 have the highest rates of markers. The rate of *uh* is slightly more affected than *um*. Even though there is an inverse tendency for the effect of naturalness rating 5 on the rate of the two markers, I cannot derive any conclusion since this is only representative of 1 conversation.

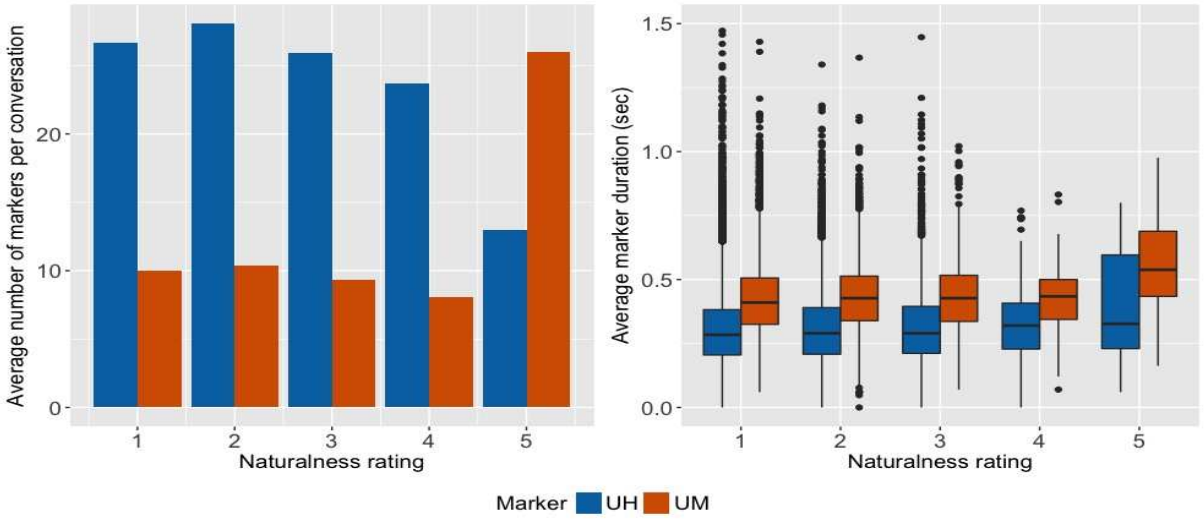

FIGURE 6.6: Effect of naturalness ratings on the average number and on the duration of tokens in *s.* per conversation in the Switchboard corpus

6.4.5 Speaker participation in Switchboard

A total of 383 speakers participated in the 1125 conversations, with a range of 1 to 25 conversations per speaker, an average of 5.9 conversations per speaker and a maximum of 25 conversations. Since this variable is speaker related, the effect of speaker participation is only investigated at the speaker level. About 1/3 of the speakers participated in 1 or 2 conversations, about 1/2 participated in 5 or less, and only 2.3% of participants participated in more than 20 conversations.

Figure 6.7 shows how speaker participation affects conversation length for both number

of words (left) and speaking time (right). The modeling function used for the regression line is loess, with a span of 0.5, and the grey area shows the 0.95 confidence interval. The largest increase in conversation length is between 2 and 7 conversations per speaker. These results show that conversation length does not vary much depending on speaker participation, especially compared the other variables used in this experiment (viz., speaker, gender and, naturalness).

Figure 6.8 plots the effect of speaker participation on the average rate (left) and duration (right) of the two markers with the same settings as Figure 6.7. Results show a slight increase in the number of tokens used by speakers when they participate in 3 to 13 conversations for *uh* and 5 to 15 for *um*, against 2 to 7 conversations for other words. Results also show no systematic trend in terms of markers' duration when speaker participation increases. Results for the production of *um* and *uh* are rather similar to results for other words, which means that they behave similarly to other words in this corpus with regards to speaker participation. Finally, it is important to note that speaker participation does not seem to affect *um* and *uh* in different ways, contrarily to other variables analyzed in this chapter.

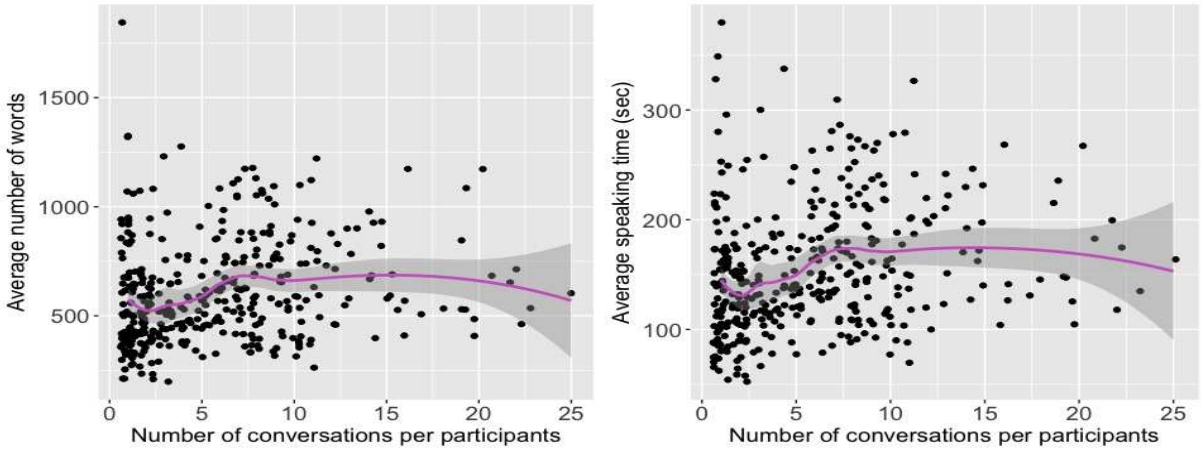

FIGURE 6.7: Effect of speaker participation in the Switchboard corpus on the average number of words and the average speaking time (in s.) per speaker

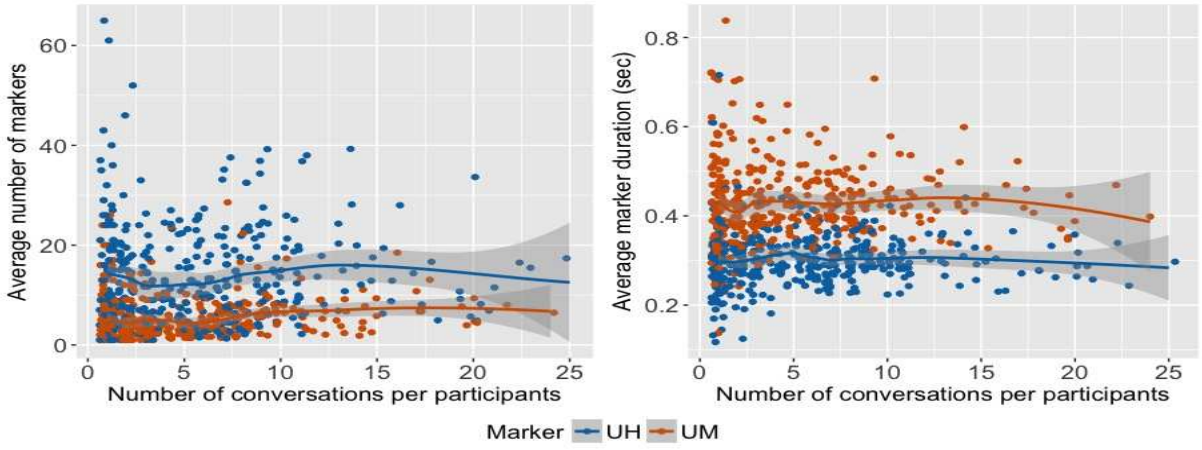

FIGURE 6.8: Effect of speaker participation in the Switchboard corpus on the average rate and duration (in s.) of *um* and *uh*

6.5 Summary and discussion

In this chapter I looked at which factors impact the rate and the duration of *um* and *uh* to find out whether they vary in systematic and similar ways. The findings from this analysis will serve as a baseline for further analyses on the two corpora. In this section, I summarize the results and compare them to see if the results carry across task (i.e., corpora).

Even though the two corpora used in this analysis consist of spontaneous spoken speech of American English, they also differ in several dimensions (dialect, time of recording, activity, topics...). These two corpora therefore only represent some aspects of spontaneous spoken speech, corresponding to different activities. The ATAROS corpus represents spontaneous speech in collaborative tasks for speakers from the Pacific Northwest who do not know each other, while Switchboard is representative of spontaneous phone conversations between strangers from various dialects.

There are 595 *ums* and 470 *uhs* in the ATAROS corpus, which respectively represent 0.92% and 0.73% of words in the corpus (see section 6.3.1). In comparison, speakers use a total of 10,784 *ums* and 30,187 *uhs* in Switchboard, which respectively represent 0.74%

and 2.07% of words (see section 6.4.2). The relative proportions of *um* vs. *uh* are inverse in the two corpora, and the difference in terms of proportions between the two markers is 7 times bigger in Switchboard than in ATAROS. The differences in the production of the two markers indicate that as far as the two corpora are concerned, there is no systematic trend in the production of the two markers.

Results on gender in ATAROS show that on average women speak for longer than men, but men use more words than women. Both genders speak for longer and with more words in the Budget Task (BT) (task that elicits more speaker involvement), and the difference between men and women is greater in the BT than in the Inventory task (see section 6.3). In Switchboard, women speak more and for longer than men, with more variability (see section 6.4.1). It is interesting to see that in both datasets women speak for longer than men because it goes against common assumptions that men speak more than women.

Duration results in both ATAROS and Switchboard show that words are significantly shorter for men than for women. However, in Switchboard, *um* and *uh* are significantly shorter for women than men, and to a larger extent for *um* than for *uh* (see section 6.4.2). Similarly, in ATAROS, *um* is also shorter for women (see section 6.3.2). This indicates that *um* and *uh* in Switchboard and *um* in ATAROS do not behave like other words with regards to duration and gender since they follow different patterns than other words. Furthermore, in ATAROS, *um* and *uh* are shorter in the Budget Task, and task has a significant effect on the duration of *um* for men. This suggests that degree of involvement affects the duration of *um* more than *uh*, and that it is greater for men than women.

Results on the rates of the two markers depending on the speaker gender in ATAROS indicate that women and men differ in the production of *um* and *uh* (see section 6.3.1). Task has a stronger effect on the rate of the two markers in men than women, which indicates that when speaker involvement changes, men are more susceptible to change their production of *um* and *uh* than women. In addition, task has a different effect on the production of markers for the two genders since men use more *ums* and *uhs* when involvement increases, whereas women use more *ums*, but use less *uhs*. Overall, across tasks, the production of *um* for women

represents 0.88% of words vs. 0.54% for *uh*, while *um* represents 0.97% of words for men, and *uh* 0.96%. In comparison, in Switchboard, women use more *ums* than men (0.84% vs. 0.62% respectively) while men use more *uhs* than women (2.76% vs. 1.53% respectively) (see section 6.4.2). Both genders use more *uhs* than *ums*, which is not surprising since speakers use on average three times more *uhs* than *ums* across Switchboard. These results suggest that there is no systematic trend in the use of the two markers across the two datasets based on speaker gender.

The results for dyad gender in ATAROS indicate that speakers use more *ums* in mixed-gender dyads than in same-gender dyads, and that speakers use more *uhs* in men-men dyads (see section 6.3.3). However, since there are fewer same-gender dyads, especially for men-men dyads, these findings need to be further investigated. Dyad gender does not have a systematic effect on the duration of the two markers. The longest *ums* on average are used between men-men dyads and the shortest ones are used between men-women dyads, while the longest *uhs* are used between women-men dyads and the shortest ones between women-women dyads. In Switchboard, dyad gender does not affect the duration of the markers, but it affects the production of *um* and *uh* in different trends (see section 6.4.3). Unlike other factors, dyad gender affects the rate of *uh* to a greater extent than the rate of *um*.

Furthermore, in Switchboard, there is no correlation between the duration of *um* and *uh* and naturalness ratings between 1 and 3 (4 and 5 are dismissed due to too few conversations). Similarly to other words, the highest rates of markers are found in conversations with a naturalness rating of 2, and as naturalness decreases (rating increases), the markers' rates decrease as well. Correspondingly to other factors, *um* is affected more than *uh* since there is a greater correlation between naturalness ratings (1-3) and the production of *um* compared to *uh*. Finally, speaker participation does not seem to have an effect on the rate or the duration of the two markers.

To conclude, most factors affect the rate and the duration of *um* and *uh*, but in different trends depending on the corpus, and therefore depending on the speech activity. However, the main conclusion is that *um* and *uh* have different distributions and duration cues, whether

in ATAROS or in Switchboard, and that most factors have a greater effect on *um* than on *uh*.

These findings, along with other studies (see section 3.1.6 in Chapter 2), indicate that the two markers be treated as two distinct entities. Furthermore, research should control for talker and listener gender, as well as the degree of involvement of the speaker, especially when looking at the duration and the rate of *um* and *uh*. In the next experiment (Chapter 7) I investigate whether they are perceived in similar ways by transcribers of the Switchboard corpus, and whether they behave like other words in terms of transcription errors.

Chapter 7

TRANSCRIPTION ERRORS OF *UM* AND *UH* IN SWITCHBOARD

7.1 Introduction

In this chapter I explore the factors that influence the production of *um* and *uh*, as well as the transcription errors made on the two markers in the Switchboard corpus. The main goals of this experiment are to analyze whether *um* and *uh* behave like other words in terms of saliency, and to find out the factors and the environments that effect transcription errors of the two markers. In order to look at transcription errors, I compare two transcript versions of the Switchboard Corpus. The original transcript (Treebank3) contains transcription errors whereas the more recent transcript (Mississippi State) is a revised and corrected version of the old one. I am interested in transcription differences concerning *um* and *uh* between the two transcript versions to find out if there are systematic transcription errors, and if there are correspondences between their perception and specific linguistic structures.

In section 7.2 I review the methodology used for this experiment regarding the data, the data processing and the analysis. In section 7.3 I present the overall distribution of the data in order to get a baseline for the rest of the analysis. Sections 7.4 and 7.5 focus on transcription errors. In section 7.4 I look at substitutions in the corpus and in section 7.5 I look at missed and hallucinated words. Finally, in section 7.6 I look at the effect of paralinguistic factors on the production of *um* and *uh*, and on the type and the number of transcription errors.

7.1.1 Goals

The goals of this experiment are to find out whether transcribers made transcription errors at the same rate for *um* and *uh*, and if *um* and *uh* were systematically substituted,

missed or hallucinated (inserted) in the original transcript depending on various variables such as transcriber, transcription difficulty, conversation duration, conversation naturalness or variables that relate to the speaker.

7.1.2 Hypotheses

Based on findings from the literature and previous experiments in the frame of this study, as well as observations of the Switchboard transcript versions, I predict the following general hypotheses to be true : H1) transcribers make less errors in transcribing lexical words than function words or discourse markers and *um* and *uh* ; H2) transcribers make more errors when transcribing *uh* than *um* ; H3) *um* and *uh* are more often missed than hallucinated in the original transcript ; H4) numerous factors have an effect on the production of *um* and *uh* and on the type and number of transcription errors made on the markers ; and finally, H5) *um* and *uh* do not behave like other words or like each other with regards to production and transcription errors. These hypotheses suggest that not only *um* and *uh* behave differently from other words in the corpus, they also suggest the two markers behave differently from each other since they co-vary with different discourse variables. If there are more errors made on the transcription of *uh* than *um*, this would also suggests that *um* is more salient than *uh*, and that *um* is more likely to play more discourse functions.

7.2 Methodology

7.2.1 Data

In this experiment I compare two transcript versions of the Switchboard corpus : the Treebank3 transcripts and the Mississippi State transcripts abbreviated *MS-State* or *MS* transcripts. Treebank3 is the original transcription that contain errors, whereas the MS-State transcripts are the more recent version with revised transcriptions of the Treebank3 transcripts. The MS-State transcripts are therefore the gold standard for this experiment, and serve as reference when comparing transcript versions. The two transcripts used in this

analysis only consist of text transcriptions of the conversations. Therefore, I use the term *Treebank3* to refer to the transcription of the conversations as opposed to the syntactic annotations. For more information about the Treebank3 and the MS-State transcripts, see section 5.4 in Chapter 5.

7.2.2 Transcription errors

Material for this experiment consists of 932 hand annotated alignments between Treebank3 and MS-State transcription versions, annotated by the MS-State team. For each file speaker and other meta data is extracted and stored in a database, which can then be linked to the alignment. The alignments contain speaker turns, speech units called *slash units*, and transcription error types. The transcription differences between the two transcript versions are referred to as *transcription errors*, classified in three categories : missed, hallucinated and substituted words.

Words not present in the original version (Treebank3) but transcribed in the corrected version (MS-State) are marked as *missed* or *M*. Example (1) illustrates the notation for missed words {{DD uh}} that signals the marker *uh* was missing from the Treebank3 transcript. Words not present in the revised version (MS-State) but present in the original version (Treebank3) are marked as *hallucinated* or *H*. Example (2) illustrates the incorrect insertion of the marker *uh* in the original transcript {{II uh }}, which means that it is not present in the corrected transcript. Words replaced by other words between the two versions are marked as *substituted* or *S* {{CC ...}}. Since the gold standard for this experiment is the MS-State version of the transcripts, I only look at substitutions where the word from MS-State is substituted but not the other way around. That is, I consider the token in the revised transcripts as the true token, and the token from the original transcripts as the incorrect one. The notation {{CC um | uh }} in example (3) means *um* is substituted by *uh*. That is, *um* is incorrectly marked as *uh* in the Treebank3 transcripts as indicated by the corrected transcripts (MsState).

(1) A.7 you {{DD uh }} you are experienced i would say
From conversation sw2837

(2) B.106 when we 've taken things to the dump just the dump {{II uh }} that is uh you know closest to us
From conversation sw3694

(3) A.61 but {{CC uh | um }} you know just like software is only given out to customers
From conversation sw2012

7.2.3 Counting conversation length

Conversation length is used as a measure of how much people spoke in a conversation. This measure does not take speaking duration or conversation duration into account, but rather the number of words spoken by each speaker in each file. This measurement is used to establish the size of the conversation and how much a speaker spoke. For instance, it is used to compute the rate of *um* and *uh*, or the rate of missed, hallucinated and substituted words, compared to how much a speaker spoke.

Conversation length is computed for speaker utterance, called slash unit (see section 5.4.3 and Meteer and Taylor (1995) for more information), overall speaking quantity for a speaker within a conversation, and for an entire conversation (how much two speakers spoke). Conversation length is computed by counting the number of words in cleaned up versions of the alignments. The cleaning up process includes removing speaker code and transcription error annotations such as those illustrated in examples (1)-(3). Transcription errors are taken into account in order to account for missed, hallucinated and substituted words. Missed items (i.e., items omitted from Treebank3) and substitutions each count as one token. Hallucinated items (i.e., items incorrectly added to Treebank3 and not present in MS-State) count as 0. Overlapping speech is counted separately for each speaker because even if it was spoken at the same time, it still counts towards the question of how much speakers spoke. For tokenization purposes, contractions such as *that 's* or *aren 't* count as

two words.

7.2.4 Word category

One of the goals of this experiment is to understand how *um* and *uh* behave not only compared to other words, but also compared to words of the same category. I use four word categories in this experiment : lexical words, function words, other words, and word fragments. I use the NLTK (Bird et al., 2009) stopword list for English to determine which words are function words. Note that word category is labeled from cleaned up conversation transcripts that do not contain parses or POS tags to provide contextual information in case of ambiguous cases. Since the primary focus of this experiment is not word category, any lexical versus function ambiguity is handled by the stopword list. That is, if a word can be either lexical or function depending on the context, the context is ignored and the word category is determined by whether the word is listed in the stopword list. In addition to the words listed in the NLTK stopwords, I added *na* from words such as *wanna* or *gonna*, the negation *nt*, and contractions containing apostrophes, to match the word tokenization used to process the data. Example (4) lists the elements added to the NLTK stopword list. The category *other* handles words that function as backchannels, filled pauses, interjections, or words that have a high token frequency and that are not typically recognized as function words. These words are not listed in the NLTK stopword list and I do not consider them as lexical words. This category contains words such as *um*, *uh*, or *um-hum* (see (5) for the full list). I also created a category called *fragments* to deal with fragmented words (e.g., *th-, becau-* or *ne-*). These words mainly behave like lexical words but some of them behave wildly differently, such as *I-*, the most frequent fragmented word with a token frequency of 2292 in the 932 transcripts. Fragmented words mostly result from a difference in transcription convention for the transcribers who did the MS-State transcript revisions. Finally, words belonging to neither category are labeled lexical words. Several samples of the data were used to verify this classification and to build the *other* category. Note that contractions count as two separate words in the NLTK stopword list. Example (6) shows how the contraction

he's is categorized : the personal pronoun *he* and the auxiliary *to be* are separately labeled function words.

(4) *nt, na, 's, n't, 're, 'm, 've, 'll, 'd, ', 'cause, 'em, don't, that's, they're, it's, isn't, aren't, didn't, you'll, doesn't, i'm, what's, hadn't, can't, haven't, you're*

(5) *um, uh, um-hum, huh, huh-uh, hum, hum-um, uh-hum, uh-huh, yeah, yep, nope, nah, oh, ah, hm, eh, ooh*

(6) He's planning = he (function) + s (function)

7.3 Overall description of the data

This section focuses on the general distribution of a few variables in the corpus, and why we need to take them into account when looking at transcription errors in the corpus, especially for *um* and *uh*.

Conversation length, as described in subsection 7.2.3 above, is a measure of how much a caller speaks. This measure essentially consists in computing the total number of words as an estimate of how much a speaker talks. The total amount of words across the 932 transcripts is 1,337,322, with an average of 1,435 words per conversation, 717 words per speaker, and 7 words per utterance (slash unit).

Table 7.1 summarizes the average, standard deviation, minimum and maximum number of words per conversation, speaker, and utterance (slash unit). All counts have high variability. Conversation length varies from 231 to 2,787 words per conversation, and from 35 to 1,901 words per speaker within a conversation. This high rate of variability indicates the importance of investigating the effect of conversation length on the rate of missed, hallucinated and substituted *ums* and *uhs* in this corpus.

TABLE 7.1: Count of words within conversation, speaker and utterance across the corpus

	Conversation	Speaker	Utterance
mean	1,434.9	717.4	7
std dev	507.2	323.5	7
min	231	35	1
max	2,787	1,901	79

The ten most common words in the data are listed in Figure 7.1. Nine words out of 10 are function word, and 1 word, *uh*, is from the *other* category. This illustrates the fact that the most common words are function words, and that *uh* is among the 10 most frequently used words. Similarly, Table 7.2 shows that function words are the most frequent words (56.5%). The counts are computed by adding the token frequency of all words within the same word category (lexical, function, other or fragment) and the percentages are computed by dividing the total number of words in each word category (Count) by the total number of words (1,337,322). Results show that fragmented words represent the smallest category in terms of token frequency, *other* words represent 6.1% of all words, and lexical words represent over a third of all words. These results also indicate that despite the fact that the category *other* is very small (15 items), it still represents 6.1% of all words, which means that words from this category have a high token frequency compared to lexical words for instance.

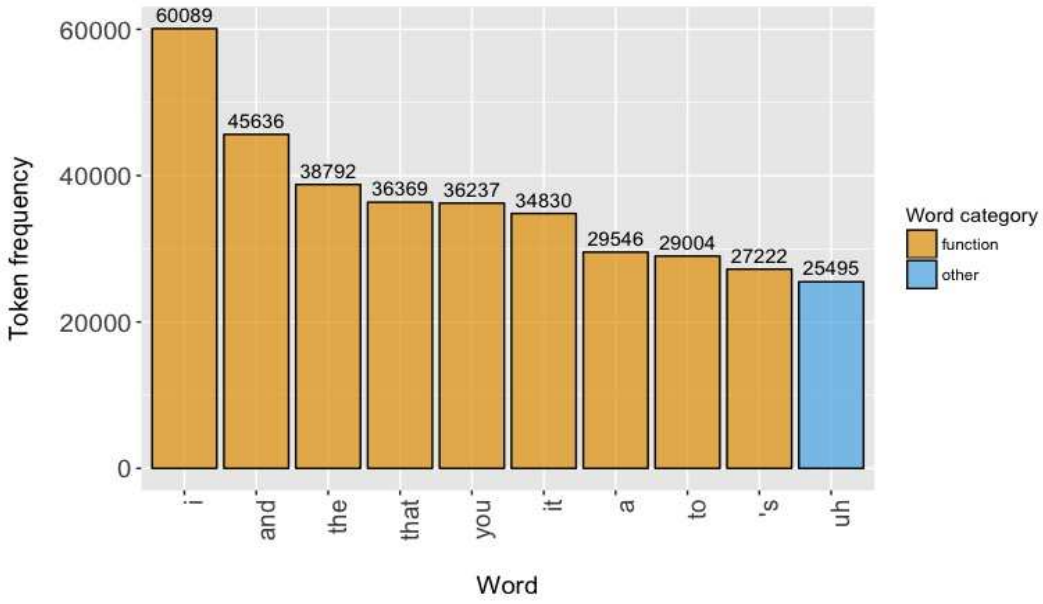

FIGURE 7.1: Top ten words with the highest token frequency

TABLE 7.2: Counts and percentages of word frequency by word category

Word Category	Count	Percentage
lexical	488,966	36.6%
function	755,666	56.5%
other	81,775	6.1%
fragment	10,915	0.8%

It is also important to take into account the transcriber variable since transcription errors are made by transcribers. In total, 31 transcribers transcribed the 932 Treebank3 conversations. Transcriber information is not available for the MS-State transcripts, which is the gold standard for the comparison. I therefore consider one transcriber for the MS-State version and I only take into account transcriber variability from the Treebank3 conversations. On

average, transcribers transcribed 30 conversations, with a wide range of variation, since they transcribed anywhere from 1 to 265 conversations. The mode is 1 transcription per transcriber, with 5 transcribers out of 31 who transcribed only 1 file. This wide range of variability indicates that the transcriber variable should be taken into account when investigating transcription errors.

Another important variable is the variability of the two markers across the corpus. Different speakers use widely different rates of *um* and *uh*. The total number of *ums* is 9,113 and the total number of *uhs* is 25,495. These totals represent the number of markers in the MS State transcripts. In other words, it excludes the number of hallucinated markers in Treebank3. The total number of *ums* is 9,166 if we include the 53 hallucinated *ums*, and the total number of *uhs* is 25,950 if we include the 455 hallucinated *uhs*. For the purpose of this analysis, I only look at the totals excluding hallucinated words. There are 2.8 more *uhs* than *ums*. Tables 7.3 and 7.4 summarize the rates of *um* and *uh* within conversation, speaker, and utterance (slash unit), to illustrate the variability of the frequency of use of the two markers. I decided to look at rates since we saw that there is a lot of variability in terms of conversation length (see Table 7.1). The rates are computed by dividing the number of *ums* and *uhs* by the number of words in the category, excluding all hallucinated words. For instance, to get the average rate of *um* used in conversations, I computed the average across all conversations of the number of *ums* used in a conversation divided by the number of words in that same conversation. Tables 7.3 and 7.4 show that the average rate of the two markers does not vary between conversation, speaker and utterance, but that the standard deviation increases as the unit decreases in size. That is, there is more variability in the production of *um* and *uh* in utterances than in speakers, and there is more variability in speakers than in conversations. The mean rates show that the rates of *uh* are about twice the rates of *um*, which is not surprising since there are 2.8 more *uhs* than *ums*. The standard deviations shows that there is slightly more variability for *uh* than *um*, which is also expected since the rates of *uh* are higher. The minimum and maximum rates show the use of *um* and *uh* varies within all three categories, and the two markers vary within a similar extent. These results

indicate that conversation, speaker, and utterance are important variables to consider in this experiment since they contain different degree in variability in terms of marker production. These results also indicate that despite the fact that the rates of *uh* are higher than for *um*, the variability of the production of the two markers is similar.

TABLE 7.3: Rates of *ums* within conversation, speaker, and utterance

UM	Conversation	Speaker	Utterance
mean	0.009	0.009	0.009
std dev	0.008	0.011	0.066
min	0	0	0
max	0.068	0.089	1

TABLE 7.4: Rates of *uhs* within conversation, speaker, and utterance

UH	Conversation	Speaker	Utterance
mean	0.018	0.018	0.017
std dev	0.011	0.015	0.073
min	0	0	0
max	0.06	0.092	1

The speaker variable is also relevant. A total of 273 speakers participated in the 932 conversations, with a frequency ranging from 1 to 23 conversations per participant, a mean of 6.8 conversations per participant, and a mode of 1 conversation per participant. About 16% of the speakers participated in only 1 conversation, about 40% participated in 5 or less conversations, about 25% participated in 10 or more conversations, and only 2.2% participated in 20 or more conversations. The heterogeneity of the number of conversations completed

by participants indicates that this factor should be taken into account in further analysis for two main reasons. We saw that speakers produce different number of *ums* and *uhs* and speakers who participate in more conversations might change speaker style over time and produce a different amount of markers than participants who only completed 1 or few conversations.

The ratio of *um* and *uh* was computed by dividing the number of *ums* and *uhs* for each speaker by how much the speaker spoke over the total number of conversations they participated in. Figure 7.2 displays the distribution of the ratios of *um* and *uh* with the smooth kernel density estimate and shows the ratio of *um* within speakers is less variable than the ratio of *uh*.

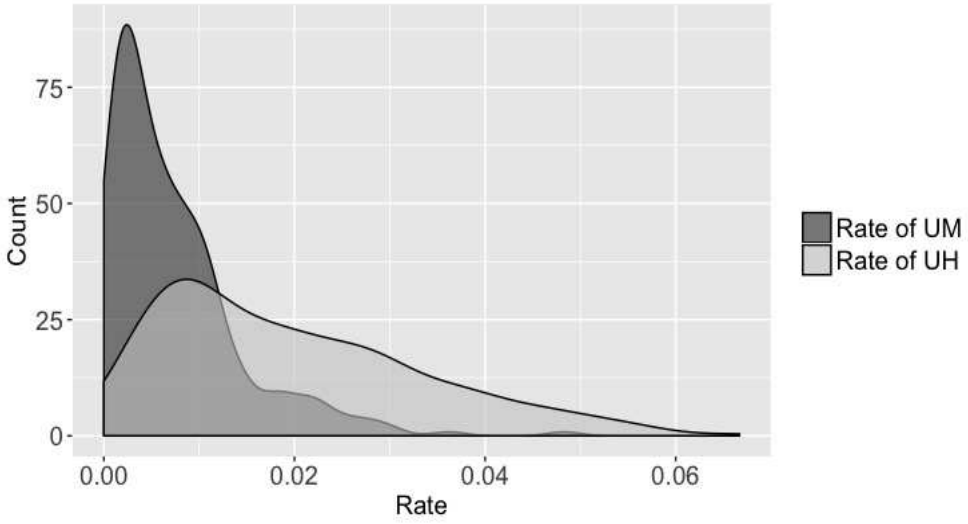

FIGURE 7.2: Speaker variability in terms of the rates of *um* and *uh*

These results show the importance of taking into account variables such as conversation length, word category, transcriber and speaker, since they affect the number of markers and transcription errors present in a conversation. Furthermore, these first results show the importance of using rates over raw counts to compensate for the variability coming from transcribers when looking at transcription errors, and for the variability coming from speakers when looking at the number of *ums* and *uhs* used in conversations.

7.4 Substitutions of um and uh

This section of the experiment focuses on the substitutions of *um* and *uh* in the MS-State transcripts by other words in the Treebank3 transcripts. That is, when a word was mistakenly transcribed as another word in the original version. As for the entire experiment, the MS-State transcripts are the reference when comparing the two transcript versions. The goal of this section is therefore to investigate any systematic patterns of substitution of the two markers between the two transcript versions to see if they tend to be substituted in any systematic way.

I first look at the substitution trends of all substituted words across the 932 transcripts to get a sense of how the two markers behave compared to other words. I also look at the most common items that substitute other words. That is, words that replaced the true words. Then I look at how *um* and *uh* behave compared to other words.

For the purpose of this experiment, I categorized three types of substitutions : 1) word fragments (e.g., *ha-, som-*) ; 2) monosyllabic function words and ; 3) multisyllabic function words as well as lexical words.

7.4.1 Results

There are 27,161 substitutions in the corpus of 6,933 kinds, across 2,401 different words that get substituted (types). Note that 4 substitutions were removed from the initial set of 27,165 substitutions, due to 4 empty words. These words result from tokenization issues when apostrophes are not followed by anything. On average, a type is replaced by 2.9 words, with a maximum of 161 words (for *'s*), and is replaced on average 11.3 times, with a maximum of 6,242 times (for *um-hum*). Across the 2,401 substituted words, 1,580 (66%) are only replaced once and 2,229 (93%) are replaced ten times or less.

There are 136 (7%) types substituted more than 10 times that are not truncated words. Of those, 95 are function words (including all forms of *be* and *have*, as well as words such as *um, uh* or *oh*), and 41 are lexical words. Only 37 types (2%) are substituted more than 100

times. Out of those, 35 are not fragment words, and 34 are function words. Only one word is lexical, *well*, although *well* is commonly used as a function word when used as a discourse marker.

Figure 7.3 plots the seven most substituted words and their word category, including the fragmented word *I-*. The top 7 most substituted words in decreasing order are *um-hum* (6242), *um* (2494), *I-* (1410), *na*, tokenized from *gonna* or *wanna*, (669), *I* (652), *uh* (641) and *'s* (580) times. The dash in *I-* indicates a partial word, which signals its transcription as an incomplete word. These occurrences most likely come from the first person personal pronoun *I* where the speaker is having some hesitation or disfluency. The next frequent partial word is *th-*, replaced 117 times by 36 words. Figure 7.3 shows that none of the seven most substituted words are lexical and that 3 of the most substituted words (*I, uh* and *'s*) are also among the 10 most frequent words (see figure 7.1). It is also important to note that 3 out of the 7 most substituted words are from the category *other*. These results show that more frequent words and words from the category *other* are more likely to be substituted than lexical words or less frequent words.

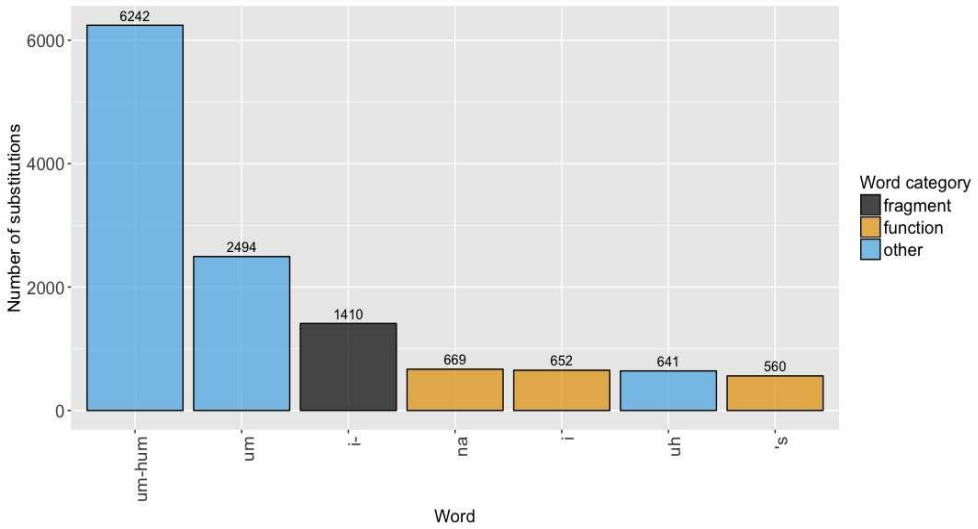

FIGURE 7.3: Seven most substituted words

There are 62 types with a substitution count greater than 50. Out of these 62 words, 39 types are function words, 12 are from the category *other*, 6 are fragmented words and 5 are lexical words (*gon, wan, well, know* and *right*). Three of the 5 lexical words, *well, right* and *know* (in the context of *you know*) are also commonly used as discourse markers, similarly to words from the category *other*. These results show that lexical words are less likely to be substituted than function words, which is expected since function words are more common (seesection 7.3). This also serves as a baseline when analyzing *um* and *uh*. For the rest of this section, I only look at function, *other*, and fragment words, which leads to a subset of the data that consists of 19,930 substitutions of 138 types.

Before looking at how the two markers are substituted by other words, I briefly look at what types of words, and how often *um* and *uh* replace other words. On average, words replace 2.9 types, with a minimum of 1 and a maximum of 100. *Um* replaces 36 types over a total of 488 substitutions, and *uh* replaces 100 types for a total of 2,941 substitutions. *Um* and *uh* therefore replace a lot more words than the average word. Figure 7.4 shows types replaced by *um* more than 10 times. All types are from the categories function and *other*, 3 of the types replaced by *um* have a nasal component, and 5 out of 6 words are monosyllabic, except for *um-hum*, which is acoustically very similar. Figure 7.5 shows the types replaced by *uh* more than 20 times. Similarly to *um*, all types are monosyllabic words, either from the categories function or *other*. *Uh* primarily replaces *um* (2,315 times). The next 3 most common words replaced by *uh* are *ah, a* and *oh*, which are acoustically similar to *uh*, especially *oh*. These results show that *um* and *uh* primarily replace words of similar category, or that are acoustically similar.

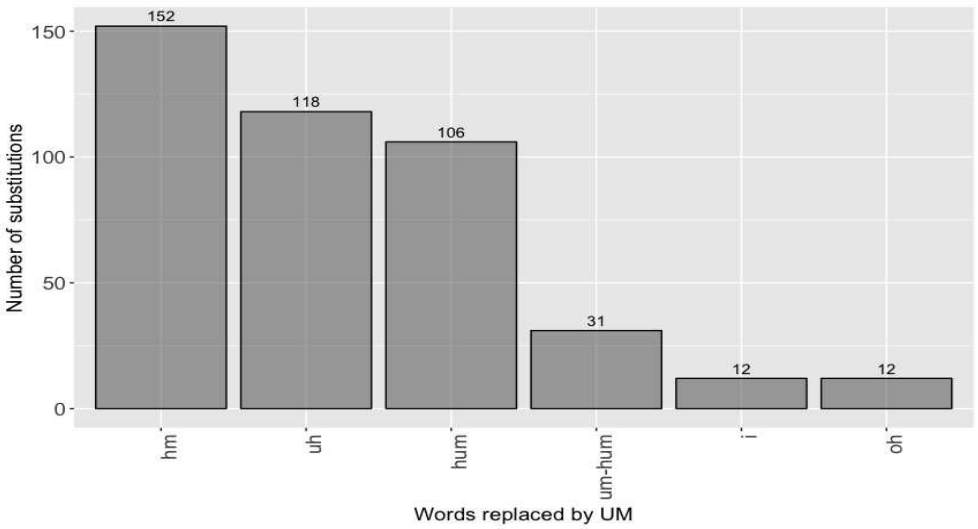

FIGURE 7.4: Words replaced by *um* more than 10 times

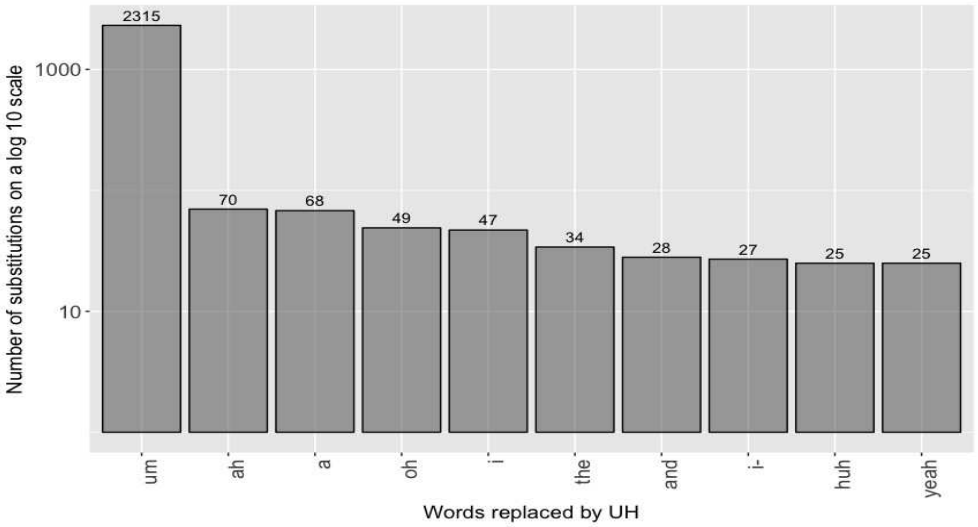

FIGURE 7.5: Words replaced by *uh* more than 20 times

In total, the marker *um* is substituted 2,494 times (12.5% of all substitutions of function words) by 44 words across the 932 transcripts. *Um* is the 14th most replaced word in terms of

how many different words replaced it, and is replaced more than twice the average function word. *Um* also has the second highest replacement count (2,494 vs. 144 on average for other function words) after *um-hum*, replaced 6,242 times.

Out of the 44 words that replaced *um* in Treebank3, *um* is substituted 7 times by 5 word fragments (*ha-, som-, n-, uh-, m-*). *Um* is also substituted by 11 words of type 3 substitution (see section 7.4, such as *working*, *around*, or *graphic*. *Um* is substituted 2,424 times by 28 words from type 2 substitution such as *that* or *oh*.

Figure 7.6 lists the count of substitution by each monosyllabic word from the categories function and other on a log 10 scale to represent the marker *uh* that replaces *um* 2,315 times in Treebank3. To summarize, the marker *um* is predominantly replaced by *uh*. Other frequent substitutions include *huh* (count = 24), *a* (count = 18), *I* (count = 14), or *oh* (count = 9). These substitution counts are not surprising since *huh*, *oh* and *a* are acoustically similar to *uh*, especially the latter. *Um* is also mistakenly replaced by *and* 7 times across the 932 transcripts, which could be due to the fact that similarly to *um*, *and* has a nasal component and is often used with *um*.

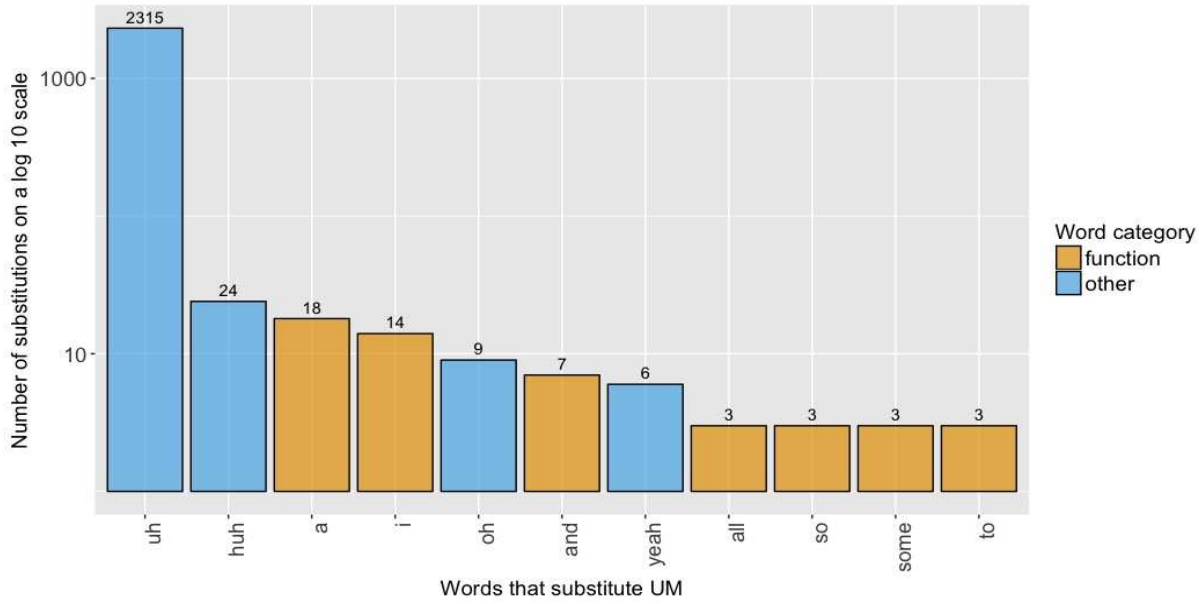

FIGURE 7.6: Monosyllabic words from the categories function and other that substitute *um* three times or more, represented on a log 10 scale

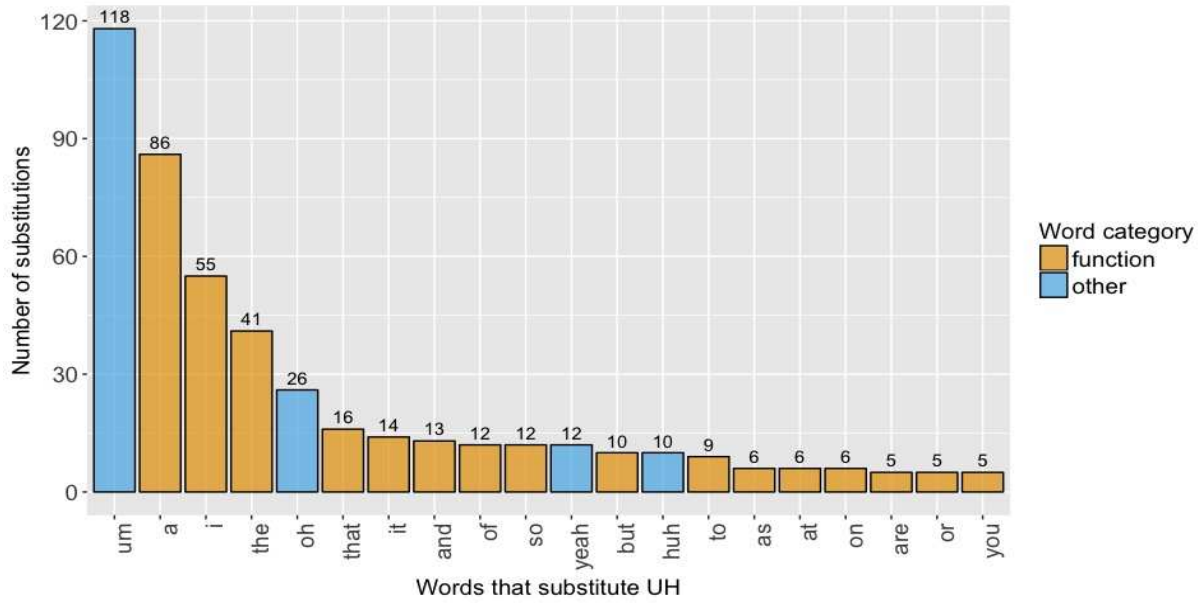

FIGURE 7.7: Monosyllabic words from the categories function and other that substitute *uh* five times or more

The marker *uh* is replaced 641 times (3.2%) by 105 words in Treebank3 across the 932 transcripts. *Uh* is the second word replaced by the highest number of words (105 vs. an average of 20 for function words) and has the fifth highest count of replacement (641 vs. 144 on average for function words) after *um-hum, um, na* and *-I*.

Out of the 105 words that replace *uh*, it is replaced 50 times by 26 word fragments and 1 punctuation mark, which is most likely an alignment issue. Similarly to *um*, it is replaced mostly once by word fragments, except for *a-*, 9 times. *Uh* is replaced 32 times by 22 words from type 3 substitution such as *husseins* or *listening*. Finally, *uh* is replaced 529 times by 55 words from the second type. Figure 7.7 plots the words that replace *uh* five times of more. Unlike *um*, replaced by *uh* 2,315 times, the marker *uh* is replaced by *um* only 118 times. These two markers therefore do not behave symmetrically in terms of substitutions. The second most common substitution of *uh* is by the indefinite article *a*, 86 times, which is expected since they are acoustically very similar.

To summarize, *um* and *uh* replace and are replaced by a lot more words that the average word. *Um* replaces 36 types and *uh* 100, against 2.9 for the average word. *Um* and *uh* are respectively replaced by 44 and 105 types, against 20 for the average word. The two markers have a substitution count higher than the average word, 2,494 and 641 respectively vs. 144 on average. *Um* is replaced a lot more than *uh*, respectively 2,494 vs. 641 times, mostly due to the fact that *uh* replaces *um* 2,315 times (93% of all substitutions for *um)*, and *uh* is replaced by more different words than *um* (105 vs. 44).

7.4.2 Discussion

Substitutions of the two markers occurring only once likely result from transcript alignment issues, especially in cases where *um* and *uh* are replaced by lexical multisyllabic words such as *graphic* which replaced *um* once or, *absolutely* and *coronary* which replaced *uh* once as well. Other substitutions might be due to transcription errors due to similarity, especially in the case of monosyllabic function words that are acoustically similar or that have similar discourse functions or environments, such as *huh*, *oh* or *a*. It is not surprising that *uh* is replaced by *a* 86 times since they are acoustically very close. It is also not surprising that *um* is replaced by *a* 18 times, since it is replaced by *uh* 2,315 times and *uh* and *a* are acoustically very similar.

The high substitution rate between *um* and *uh* is very interesting, especially the vast disparity in terms of substitution proportion between the two markers. The substitution rate of *um* by *uh* represents 93% of all substitutions for *um*, whereas the substitution of *uh* by *um* represents only 18% of all substitutions for the marker *uh*. This proportion gap signals that is is more likely for transcribers to put *uh* in place of *um* than the opposite. What is also important to note is that when we subtract the substitutions of *um* by *uh* there are only 179 substitutions of *um* remaining, which means that apart from the substitution of *um* by *uh*, *um* is less often substituted than *uh* (179 vs. 641). This might be due to various factors. For instance, the duration of *um* might have an effect on the likelihood to perceive the nasality of the marker. The shorter the marker is, the more likely it might be replaced with *uh*.

7.5 Missed and hallucinated ums and uhs

This section investigates the rates of missed and hallucinated *ums* and *uhs* in the Treebank3 transcript version, compared to the MS-State version, which is the revised version of the transcripts and therefore the reference for comparison. The goals of this section are to investigate whether *um* and *uh* are more often missed or hallucinated than other words in the corpus, and whether there are systematic patterns factoring in these transcription errors for the two markers. In this chapter, I focus on the question of whether the rates of missed and hallucinated words depend on specific variables such as transcriber, speaker, naturalness and difficulty ratings.

7.5.1 Results on overall distribution and variation of the data

I first looked at any kind of missed and hallucinated words across the 31 transcribers and the 932 conversations in order to get a baseline of the overall transcription error rates across the corpus. These baseline rates are then used to compute the rates of transcription errors for *um* and *uh* compared to the overall rate of errors, and compared to other words or words from the same word category (see section 7.2.4).

The total number of missed words across the corpus is 27,159, and 6,657 for hallucinated words. These raw numbers show that there are 4 times more missed words than hallucinated words, which means that transcribers tend to miss words 4 times more often than inserting non-existent words in Treebank3. Tables 7.5 and 7.6 list the total number of missed (M) and hallucinated (H) words for the two markers, as well as the rates of missed and hallucinated markers compared to the total number of missed or hallucinated words (*Error rate*), and compared to the total number of *ums* or *uhs*.

The rates show that *uh* is missed more than *um*. Missed *uhs* represent 10% of all missed words, versus 2.2% for *um*, and 6.5% of *ums* are missed versus 10.3% of *uhs*. Similarly, *uh* is more often hallucinated than *um*. The rate of hallucinated *uhs* represents 10% of all hallucinated words whereas *um* only represents 0.8% of all hallucinated words, and 10.3% of

uhs are hallucinated against 0.7% for *ums*. These rates are important because they provide information on the overall trends of transcription errors for *um* and *uh* but they do not inform us on how they perform compared to other words, and compared to words of the same category. This issue is addressed in the following subsection.

TABLE 7.5: Total and rates of missed (M) and hallucinated (H) *ums* computed over the number of transcription errors of the same type (Error rate), and over the total number of *ums* (*Um* rate)

Marker	Transcription error	Total count	Error rate	*Um* rate
um	M	598	0.022	0.065
um	H	53	0.008	0.006

TABLE 7.6: Total and rates of missed (M) and hallucinated (H) *uhs* computed over the number of transcription errors of the same type (Error rate), and over the total number of *uhs* (*Uh* rate)

Marker	Transcription error	Total count	Error rate	*Uh* rate
uh	M	2,676	0.1	0.1
uh	H	455	0.068	0.018

7.5.2 Errors of um and uh compared to individual words and word category

In this section, I compare the proportions of missed and hallucinated markers to the error rates of other words, and their word category (see lemma 7.2.4). The goal of this section is to better understand how *um* and *uh* are missed or hallucinated compared to other words, and compared to words of the same category.

In total, there are 1,337 missed individual words (types) for a total of 27,159 missed tokens (all occurrences). The number of times words are missed ranges from 1 to 2,676. There is a total of 552 hallucinated types out of 6,657 hallucinated tokens, with the number of substitutions per type ranging from 1 to 774. These numbers show that there is a lot more variation among the missed words than among the hallucinated ones, which is not surprising since there are four times more missed words than hallucinated words.

There are 488,966 lexical token in total in the corpus, 755,666 function tokens, 81,775 tokens from the *other* category, and 10,915 fragment tokens. Function words are unsurprisingly the most frequent, and fragment words are the least frequent. Table 7.7 lists the proportions of words within error type (missed vs. hallucinated) for each word category (lexical, function, *other*, and fragment). The proportions show that function words represent the largest part of missed and hallucinated words. Function words represent about half of missed words, and almost two thirds of hallucinated words. These proportions show that function words tend to be hallucinated more than missed, compared to other words. Lexical words represent similar proportions of missed and hallucinated words (13.4% vs. 13.6%). Words from the *other* category represent a larger proportion of missed words than hallucinated words (21.7% vs. 16.2%) and fragment words also represent a wider part of missed words than of hallucinated words (14.5% vs. 6.9%). These results suggest that *other* and fragment words tend to be missed more than hallucinated.

Table 7.7: Proportions of word category by error type

Word Category	Missed	Hallucinated
lexical	13.4%	13.6%
function	50.4%	63.3%
other	21.7%	16.2%
fragment	14.5%	6.9%

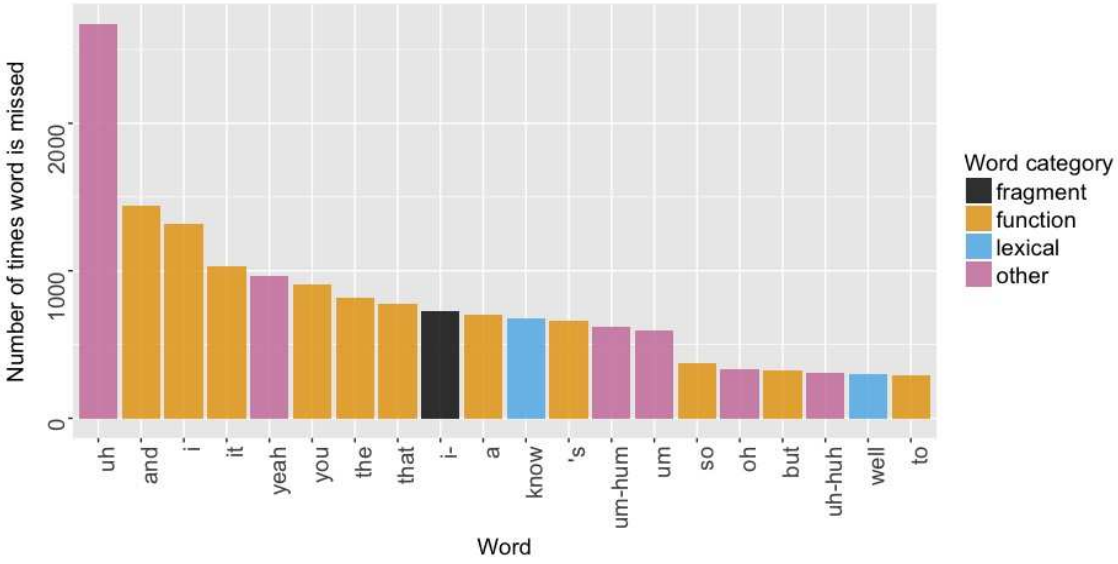

FIGURE 7.8: 20 most missed types in decreasing order and grouped by word category

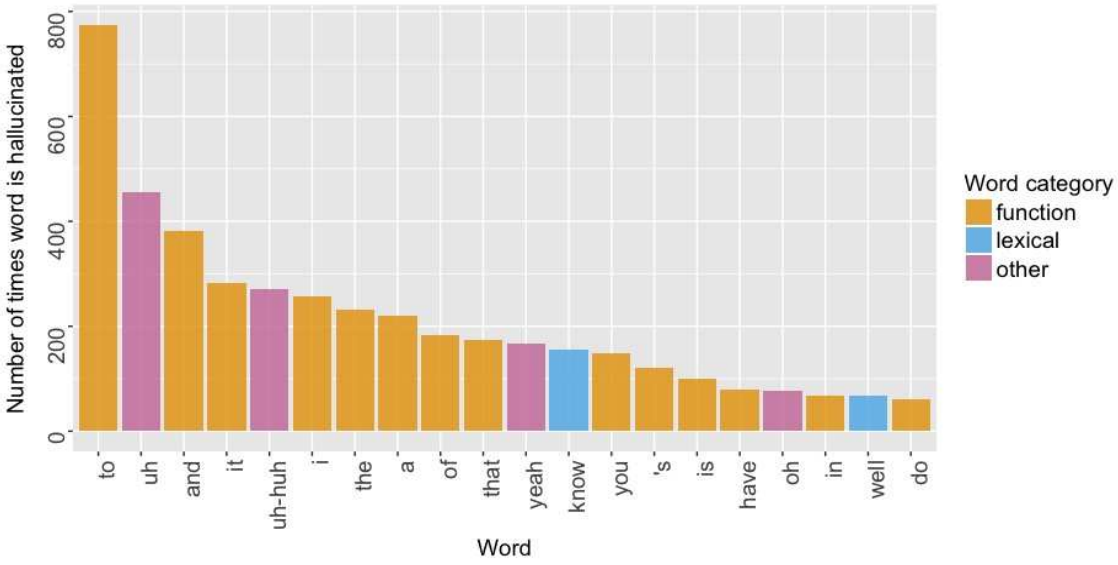

FIGURE 7.9: 20 most hallucinated types in decreasing order and grouped by word category

Figures 7.8 and 7.9 respectively list the 20 most missed and hallucinated types and their word category. These figures show that *uh* is more often missed (2,676 times) than *um* (598 times) and that *uh* is the most missed type, 1,236 times more often than the second most missed type *and*. *Uh* is also more often hallucinated than *um*, respectively 455 vs. 53 times, and is the second most hallucinated type, after *to* (hallucinated 774 times), and before *and* (hallucinated 381 times). In contrast, *um* is the 26th most hallucinated type and the 14th most missed type.

The number of lexical words is far inferior to the number of function and other words when the error count is equal to or greater than 100. Figures 7.8 and 7.9 show that among the 20 most missed and hallucinated types, only two word types are lexical : *know* and *well*. These two types are categorized by the NLTK stopword list as lexical words. However, in several contexts *know* and *well* are used as discourse markers, especially when *know* is used in the context of *you know*. *Well* and *know* are therefore not purely lexical items since they can also be used as function words. It is therefore not surprising that *know* and *well* are present in the top 20 most missed and hallucinated words, since they are often used like words that behave similarly.

Figure 7.10 plots the percentages of missed (left) and hallucinated (right) words by their percent total, for each word category (fragment, function, lexical, and other). Note that there is only one data point for fragment words. This is because all fragment words are collapsed under the same category here, because they mostly are lexical words, actual fragment words, or words issued from changes in transcription conventions. The plots for both missed words and hallucinated words shows that function words and *other* words have more transcription errors that lexical words, and *other* words have more transcription errors than function words, especially in the case of missed words. Figure 7.11 only plots function and *other* words and shows that proportionally to their percent total, *other* words are more missed than function words, and *uh* is proportionally more missed and hallucinated than *um*. It is also interesting to note that compared to its percent total, the word *yes* is less hallucinated than words from the same category, and is among the lowest ratios of percent total and percent missed. This

may be due to the fact that the word *yes* is a difficult word to assign to a word category, and that it behaves more like lexical words than like *other* words, unlike words such as *yep*, *nope*, or *yeah*. These results indicate that *other* words behave differently from words of different word categories, and that they are more likely to be missed than other frequent words, such as function words.

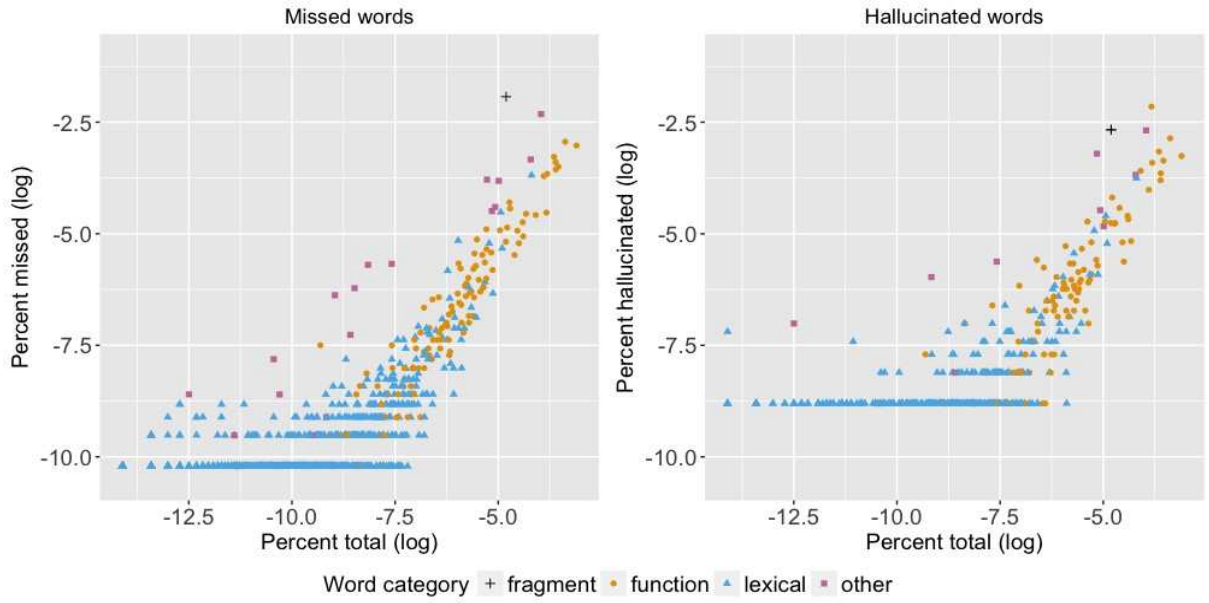

FIGURE 7.10: Log of percent total by log percent error (missed on the left, hallucinated on the right) for the four word categories (fragment, function, lexical, and other)

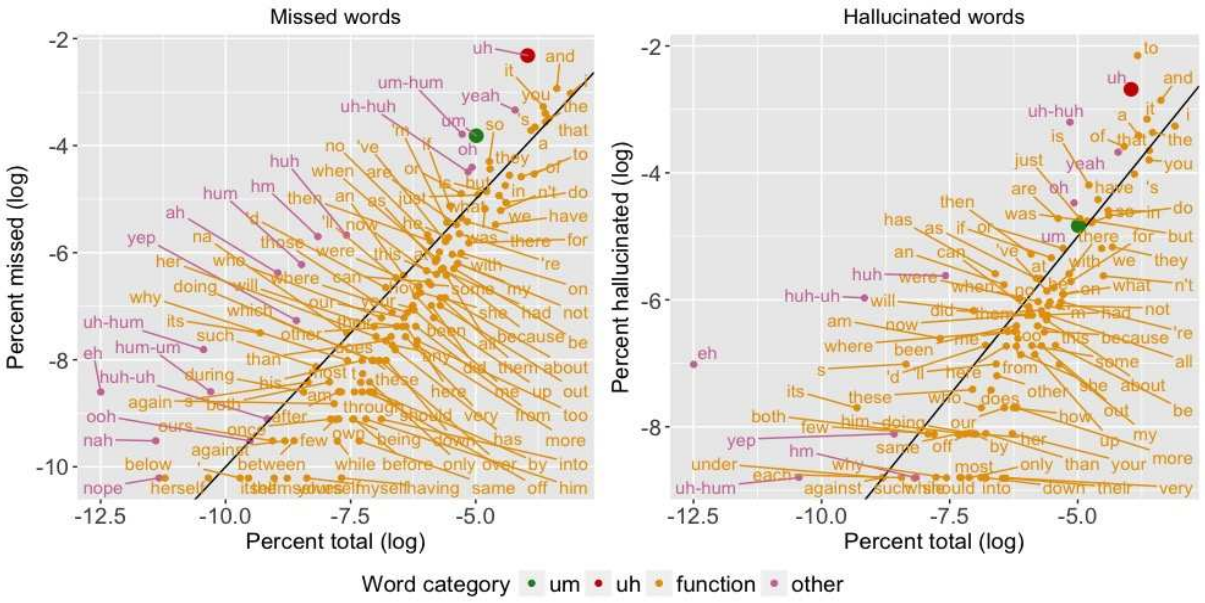

FIGURE 7.11: Log of percent total by log percent error (missed on the left, hallucinated on the right) for function and *other* words, with *um* indicated by the green dot and *uh* indicated by the red dot

7.5.3 Summary

The results from this section show that transcribers are more likely to miss words than to hallucinate them. The marker *uh* represents a greater proportion of error types (misses and hallucinations), and greater missed and hallucinated rates than *um*. Results on word category indicate that proportionally to their percent total, function words are more likely to be hallucinated and missed by transcribers than lexical words, and words from the *other* category are as likely to be hallucinated and more likely to be missed than function words. Words that have a high error rate are more likely to be function or *other* words, than lexical words. And finally, *uh* is more likely to be missed or hallucinated proportionally to its token frequency than *um*, especially compared to words from the same category. Based on results from Chapter 6 that indicate *um* and *uh* have different distributions depending

on paralinguistic factors, the next section investigates the effects on the rate of transcription errors that pertain to the transcriptions and to the speakers.

7.6 Paralinguistic variables

7.6.1 Results on transcriber effects

In this section I investigate effects pertaining to the transcriber on the overall number of missed and hallucinated words, as well as for the two markers *um* and *uh*. Variables related to the transcriber include the transcriber's identity, the number of files transcribed by the transcriber, the date on which the files were transcribed and, ratings annotated by the transcriber on the naturalness of the conversation and the difficulty to transcribe it.

Effect of number of transcriptions per transcriber on transcription error rate

As previously mentioned in subsection 7.3, 31 transcribers transcribed between 1 and 265 Treebank3 conversations. Figure 7.12 shows the variation in the rate of missed, hallucinated, and substituted words for each transcriber. The error rates are computed by dividing the number of errors in a conversation by the conversation length. Figure 7.12 shows the violin plots which combine box plots and kernel density plots to show the distribution shape of the data. The wider areas represent higher probabilities for members of the population to take the given values and narrower areas represent smaller probabilities. The empty circles at the bottom and the top of the violins represent outliers, and dots represent small amounts of data. These results show that the number of transcription errors vary across and within transcribers, which is likely caused by the fact that the number of files transcribed by transcribers is not homogeneous. Several transcribers transcribed only one or a few files while others transcribed up to 265 files. Even though the transcription error rate is normalized by the conversation length, it does not account for the number of transcriptions the transcriber completed. The plots also show that the same transcribers do not systematically make equivalent amounts of errors across transcription error types. For instance, transcriber HJR made

a lot of substitution errors compared to other transcribers, but missed less words than other transcribers. These results indicate the importance of taking the number of conversations transcribed by each transcriber into account since it can affect the analysis on transcription errors.

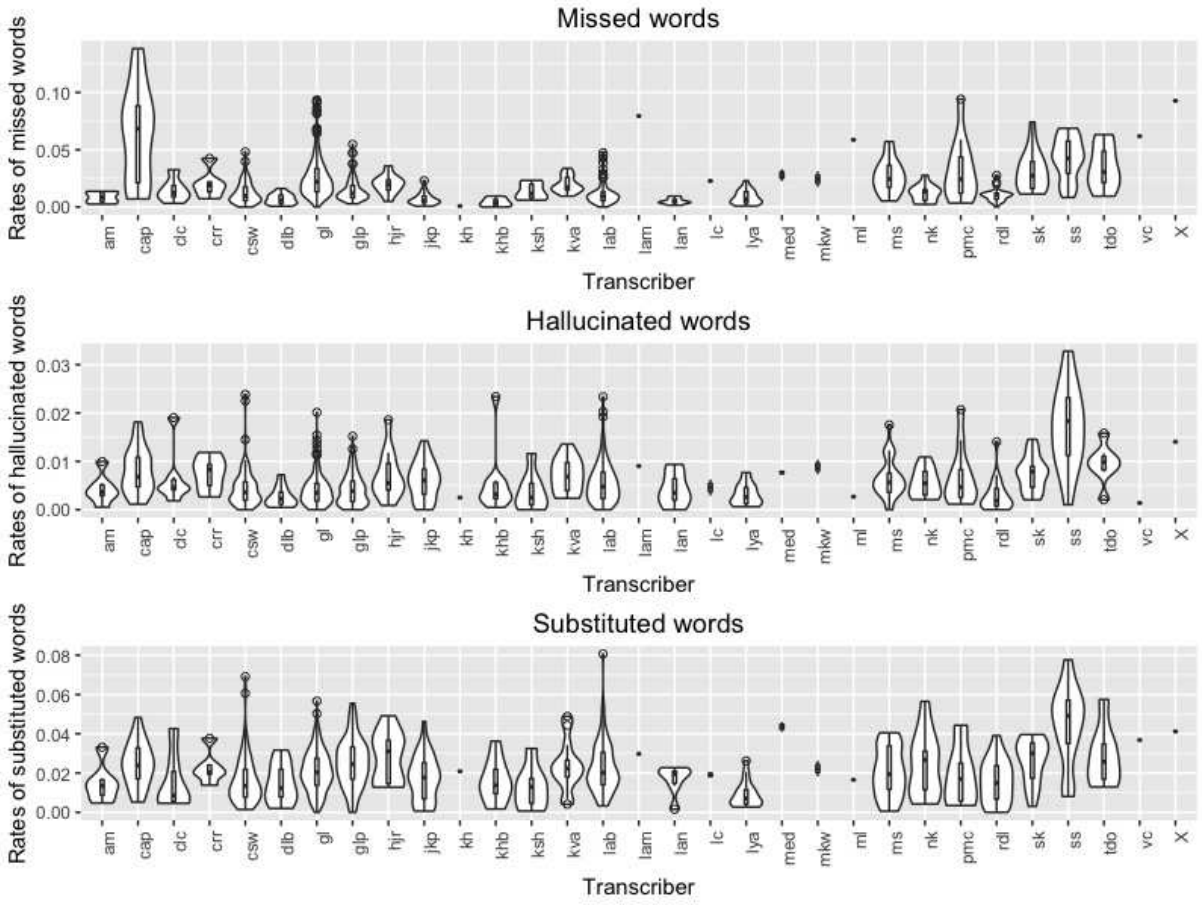

FIGURE 7.12: Violin plots (combined box plots and kernel density plots) of the rates of missed, hallucinated, and substituted words within transcriber and across conversations. Violin plots show the distribution shape of the data, wider areas represent higher probabilities for the values while narrower areas represent smaller probabilities, empty circles at the top and at the bottom of the violins represent outliers, and dots represent small amounts of data

Figure 7.13 plots the effect of the number of conversations transcribed by each transcriber on the rates of missed, hallucinated, and substituted words by each transcriber. Each dot represents a transcriber, and the transcription error rates are computed by dividing the number of each transcription error type in the conversation(s) completed by the transcriber by the total number of words spoken in the conversation(s). Figure 7.14 illustrates the same principle for missed, hallucinated, and substituted *ums* and *uhs*.

Results presented in Figures 7.13 and 7.14 show a lot of variability in the number of transcription errors made by transcribers who only completed a few conversations. We cannot get the same information for transcribers who did a high number of transcriptions because there are not enough of them. The two figures also show that transcribers who did more transcriptions tend to make less transcription errors on *um* and *uh* than on other words, as indicated by the rates. That is, transcribers who transcribed more conversations seem less likely to miss, hallucinate, or substitute *um* and *uh*. It is also interesting to note that there is an outlier transcriber who transcribed 170 conversations and who has a very high rate of substitutions, especially for the marker *um*. This transcriber has the highest rate of substitutions, 0.0057 for *um* vs. 0.001 for the average transcriber. This means that if we exclude this outlier transcriber for substitutions, the effect of transcription number per transcriber is similar for *um* and *uh* for the three types of transcription errors.

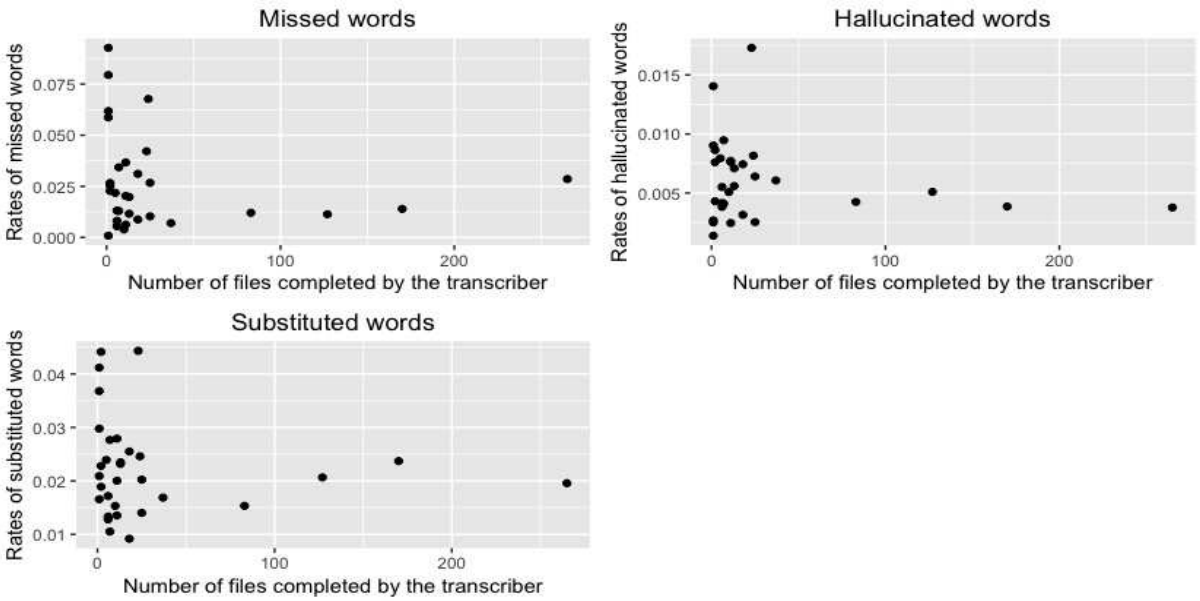

FIGURE 7.13: Rates of missed, hallucinated, and substituted words depending on the number of conversations completed by each transcriber, indicated by a filled dot

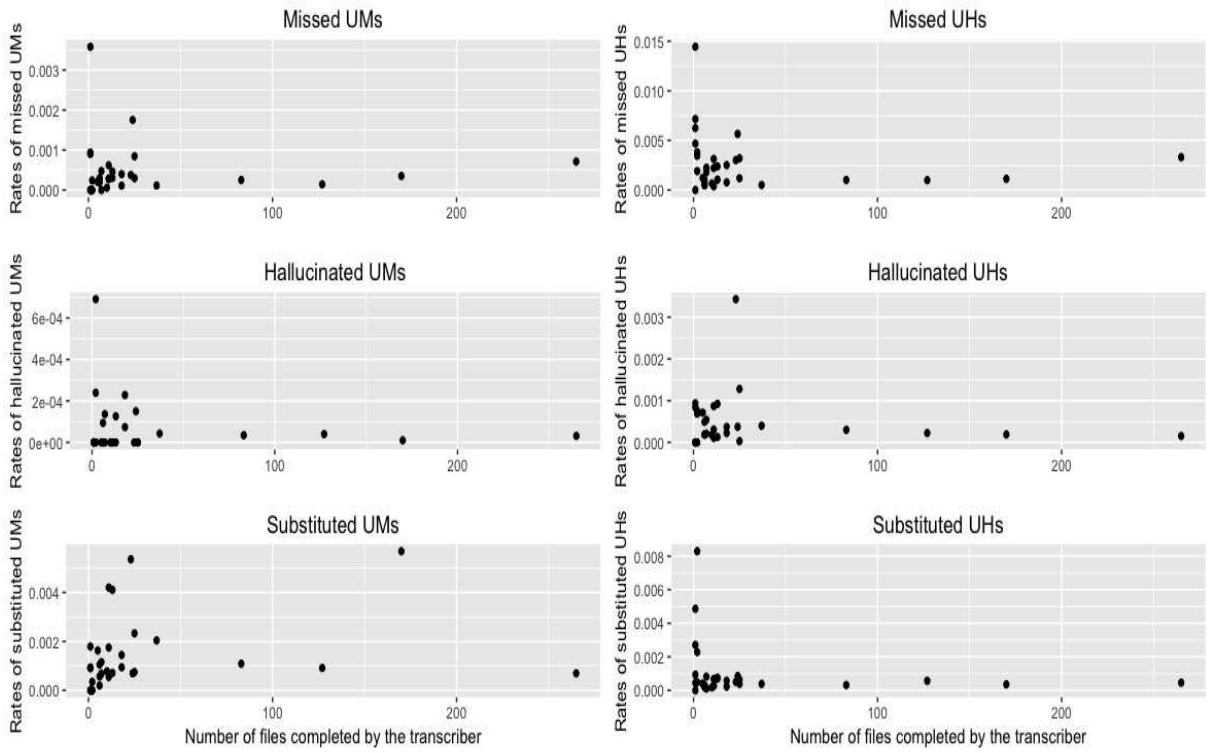

FIGURE 7.14: Rates of missed, hallucinated, and substituted *ums* and *uhs* depending on the number of conversations completed by each transcriber, indicated by a filled dot

Effect of transcription date on transcription error rate

Transcription date is another relevant factor to take into account when looking at transcription error rate. In other words, I am interested in finding out whether there is a fatigue or a learning effect as transcribers do more transcriptions.

Figure 7.15 illustrates the effect of transcription date on the number of transcription errors for two transcribers, CSW and JKP, who respectively completed 83 and 37 transcriptions on 38 and 11 different dates. Results show the error rates for missed, hallucinated, and substituted *ums* and *uhs*. Each point on the plot represents an error rate. This figure illustrates how transcription date has different effects for each transcriber, marker, and error type (missed, hallucinated and substituted markers). In addition to having various effects,

all transcribers who completed several transcriptions have different number of transcriptions completed on different dates, which leads to not enough data points to perform a statistical analysis of the effect of transcription date on the rates of transcription errors of *um* and *uh*.

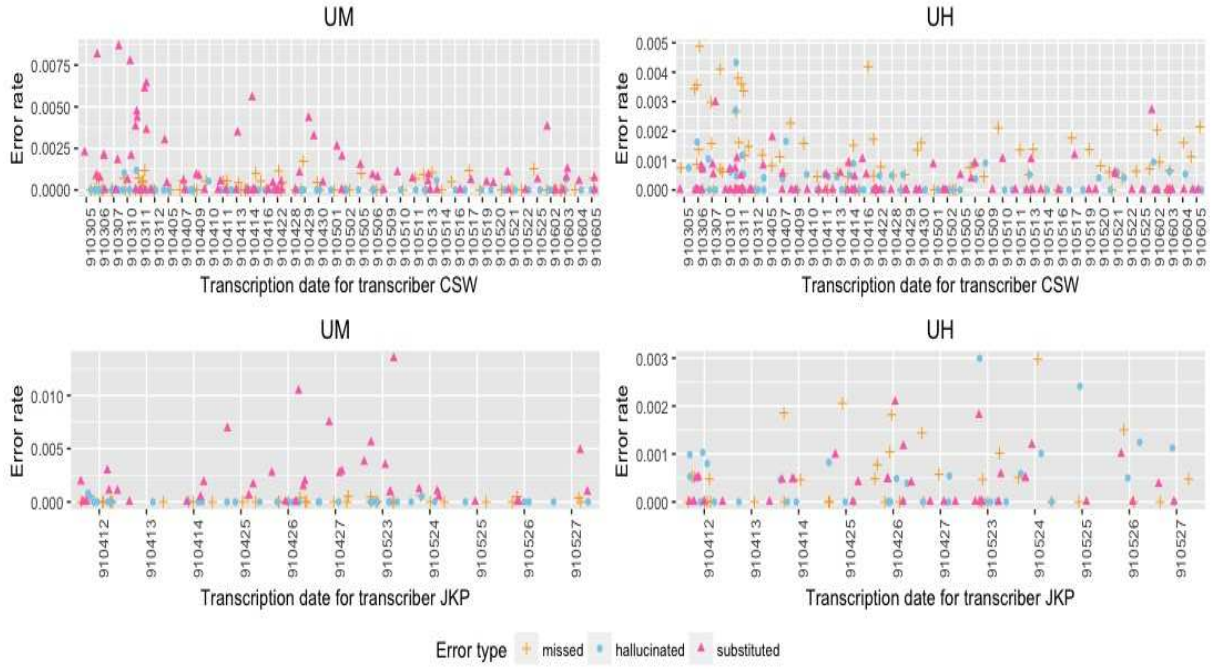

FIGURE 7.15: Example of error rate for *um* and *uh* by date of transcription (YYMMDD) for transcribers CSW who transcribed 83 conversations and JKP who transcribed 37 conversations

Effect of difficulty and naturalness ratings on transcription error rate

In this section, I am especially interested in testing two ratings assigned by transcribers : naturalness and difficulty. Difficulty is rated on a scale 1-5 to estimate the difficulty of transcribing a conversation. A low rating (1) means the conversation is easy to transcribe and a high rating (5) means the task is difficult (Godfrey and Holliman, 1993). I hypothesize that when transcribers rate a conversation as difficult, they are more likely to make transcription

errors. That is, transcribers are more likely to miss, hallucinate, or substitute *um* and *uh* than when a conversation is rated 'easy' to transcribe. Naturalness is also on a scale 1-5 and indicates whether the transcriber finds the conversation natural between the two speakers (Godfrey and Holliman, 1993). A low naturalness score indicates that a conversation sounds natural and a high score means that it sounds artificial (see section 5.4.1 in Chapter 5). I hypothesize that in more natural sounding conversations transcribers tend to miss more *ums* and *uhs* and that when they are less natural they tend to hallucinate and substitute more markers. The reasoning behind this hypothesis is that in natural conversations speakers/listeners tend to pay less attention to disfluencies and are therefore more likely to miss them. On the contrary, *um* and *uh* are more likely to be more salient in less natural conversations than in a natural context, and therefore less likely to be missed.

To test these two hypotheses I use linear mixed effect models with random intercept. The random effects are the conversation and the speaker, and the fixed effects are the naturalness and difficulty ratings. The dependent variable is the number of transcription errors, missed, hallucinated, or substituted markers, analyzed separately. Model significance is computed using the Likelihood Ratio Test to compare the full model to reduced models. The analysis is conducted on 9 data sets : missed, hallucinated, and substituted words, *ums* and *uhs*. Table 7.8 summarizes the results for all words and table 7.9 summarizes the results for *ums* and *uhs*.

TABLE 7.8: Result summary on the effect of the difficulty and naturalness ratings on missed, hallucinated, and substituted words

Data	Missed words	Hallucinated words	Substituted words
Naturalness	$p < 0.001$	$p < 0.01$	$p > 0.05$
Effect	decreasing	decreasing	N/A
Difficulty	$p < 0.01$	$p < 0.05$	$p > 0.05$
Effect	increasing	increasing	N/A

TABLE 7.9: Result summary on the effect of the difficulty and naturalness ratings on missed, hallucinated, and substituted *ums* and *uhs*

Data	M *ums*	H *ums*	S *ums*	M *uhs*	H *uhs*	S *uhs*
Naturalness	$p > 0.05$	$p > 0.05$	$p < 0.001$	$p < 0.05$	$p > 0.05$	$p > 0.05$
Effect	N/A	N/A	increasing	decreasing	N/A	N/A
Difficulty	$p > 0.05$	$p > 0.05$	$p < 0.01$	$p > 0.05$	$p > 0.05$	$p > 0.05$
Effect	N/A	N/A	increasing	N/A	N/A	N/A

Results in table 7.8 show that difficulty has a significant effect on the rates of missed ($\chi^2(1) = 8.8$, $p < 0.01$) and hallucinated words ($\chi^2(1) = 6.4$, $p < 0.05$), increasing missed

words by 2.120e-03 and hallucinated words by 4.256e-04. Naturalness also has a significant effect on the rates of missed words ($\chi^2(1) = 14$, $p < 0.001$), as well as on hallucinated words ($\chi^2(1) = 7.8$, $p < 0.01$), and has a decreasing effect of -3.504e-03 for missed words and -6.160e-04 for hallucinated words. Results show that difficulty and naturalness do not have a significant effect on the rates of substituted words.

Table 7.9 summarizes the results on the effect of difficulty and naturalness ratings on the rates of *um* and *uh*. Results indicate that difficulty and naturalness ratings do not have a significant effect on the rates of missed or hallucinated *ums*. However, difficulty and naturalness have a significant effect on the rates of substitutions of *um*, ($\chi^2(1) = 10.8$, $p < 0.01$) for difficulty, increasing the rates of 4.278e-04, and ($\chi^2(1) = 71.6$, $p < 0.001$) for naturalness, increasing the rates by 1.454e-03. Finally, only naturalness has a significant effect on the number of missed *uhs* ($\chi^2(1) = 5.2$, $p < 0.05$), decreasing the number of missed *uhs* by -2.856e-04.

7.6.2 Summary and discussion

It is important to note that the differences between the levels of naturalness and difficulty are small (e.g., 2.120e-03). These numbers indicate variation among the rates, which are small values to begin with. It is therefore important to keep in mind that some of the results show a significant effect, and that the effect is small.

Um and *uh* behave similarly to other words and to each other with regards to the effect of the number of transcriptions done by the transcribers on the rates of missed, hallucinated, and substituted words. However, *um* and *uh* behave somewhat differently from other words with regards to the effect of difficulty and naturalness ratings on the transcription error rates. Difficulty and naturalness have a significant effect on the rates of missed and hallucinated words, but not on substituted words. On the contrary, naturalness and difficulty have a significant effect on substituted *ums*, but not on missed or hallucinated *ums*. Difficulty and naturalness do not have a significant effect on substituted *uhs*, but naturalness has a significant effect on the rates of missed *uhs*. These results show that difficulty and naturalness

affect the two markers differently in terms of transcription error rate. In addition, naturalness has a decreasing effect on the rates of missed and hallucinated words, and on the rates of missed *uhs*, but it has an increasing effect on the rates of substituted *ums*. Since the effect is different for *um* because it only affects substitution rate, I cannot compare *um* to other words and *uh* in this regard. However, the results show that naturalness has a similar effect on missed *uhs* than on other words. Finally, difficulty has an increasing effect on missed and hallucinated words, and on substituted *ums*.

These results show that with regards to difficulty and naturalness, *um* and uh somewhat behave differently from other words, and from each other. To go back to the initial predictions, substituted *ums* behave as predicted by the hypothesis with regards to difficulty. That is, as difficulty increases, transcribers are more likely to make transcription errors. Predictions for naturalness are that transcribers are more likely to miss *um* and *uh* in more natural conversations, while they are more likely to hallucinate or substitute the markers in less natural conversations. It is important to remember that naturalness ratings are inverse, which means that a score of 1 indicates the highest degree of naturalness. The results indicate that naturalness has an increasing effect on substituted *ums*. That is, in less natural sounding conversations, transcribers are more likely to substitute *um* for another word. Finally, results indicate that naturalness has a decreasing effect on missed *ums*. Again, since the ratings are inverse, this means that in less natural conversations, transcribers are less likely to miss *uhs*, which goes with the predictions that people are more likely to miss *um* and *uh* in more natural sounding conversations.

7.6.3 Results on speaker effects

The first goal of this section is to investigate the effect of speaker (referred to as speaker ID or caller), gender, age, dialect, and education level on the production of *um* and *uh*. The second goal is to find out whether participants who participated in the study more than other participants produce more or less markers. The data for this section consists of 932 conversations and 273 speakers who participated in 1-23 conversations with an average of 6.8

conversations per caller and a mode of 1. Note that there are five levels of education level, less than high school (0), less than college (1), college (2), more than college (3), and unknown (9). Since the latter category only contains 4 observations it is excluded from the analysis. See section 5.4.2 for more information on the participants of the Switchboard corpus.

Results (summarized in Table 7.10) were submitted to linear mixed effect models with random intercept to investigate the effect of speaker ID, gender, age, dialect and education level, on the rates of *um* and *uh* produced by speakers. The random effect is the conversation, and the fixed effects are a combination of speaker ID and one of the four remaining factors e.g. speaker ID + age, speaker ID + gender, or speaker ID by itself. The dependent variable is the rate of *um* or *uh*.

The analysis shows the speaker variable has a significant effect on the production of *um* ($\chi^2(1) = 19.7$, $p < 0.001$) and *uh* ($\chi^2(1) = 4.9$, $p < 0.05$), which means that different speakers produce significantly different numbers of markers. Results show that gender also has a significant effect on the production of *um* ($\chi^2(1) = 58.7$, $p < 0.001$) and *uh* ($\chi^2(1) = 343$, $p < 0.001$). Men use less *ums* and more *uhs* than women (see Figure 7.17). The South Midland dialect is the only dialect that has a significant effect on the production of *um* ($\chi^2(1) = 40.6$, $p < 0.001$). The dialects that have a significant effect on the production of *uh* ($\chi^2(1) = 23$, $p < 0.01$) are the Northern, Southern, North Midland, and NYC dialects. The fact that not all dialects have a significant effect on the production of the two markers is not surprising since Godfrey and Holliman (1993) mentions a priori classification of the participants' dialect having limited effect on predicting speech patterns. Participant's age also has a significant effect on the production of *um* ($\chi^2(1) = 111.6$, $p < 0.001$) and *uh* ($\chi^2(1) = 83.7$, $p < 0.001$). Older speakers use less *ums* and more *uhs* than younger speakers as illustrated in Figure 7.16. Finally, results on education show that education level only has a significant effect on the production of *uh* ($\chi^2(1) = 4.7$, $p < 0.05$), mainly driven by the group of speakers who stopped their education prior to high school (0). Results illustrated in Figure 7.18 show that the production of *uh* decreases between the groups 0 through 2, which means that people who went to college use less *uhs* than people who did not go to

high school or college.

TABLE 7.10: Result summary on how speaker variables affect the production of *um* and *uh* - grey cells represent signicant effect of the variable on the production of the marker

Fixed effects	ID	ID + gender	ID + age	ID + dialect	ID + education
Rate of *um*	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p > 0.05$
Effect on *um*	random	W > M	increasing	random	random
Rates of *uh*	$p < 0.05$	$p < 0.001$	$p < 0.001$	$p < 0.01$	$p < 0.05$
Effect on *uh*	random	W < M	decreasing	random	decreasing

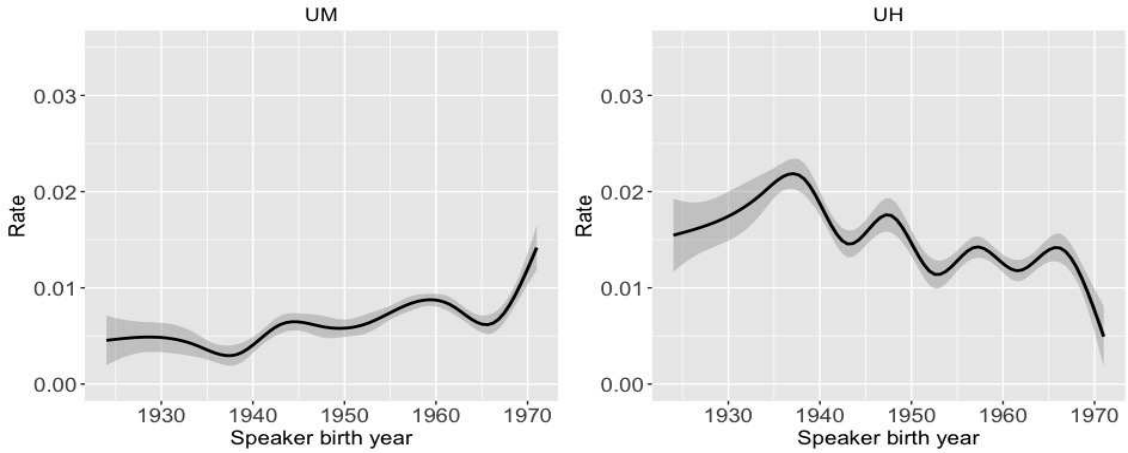

FIGURE 7.16: Effect of speaker age (plotted in birth year on the x axis) on the production rates of *um* and *uh* - smoothing method used GAM (generalized additive model) with grey shaded area representing the 95% confidence interval of the smoothing method

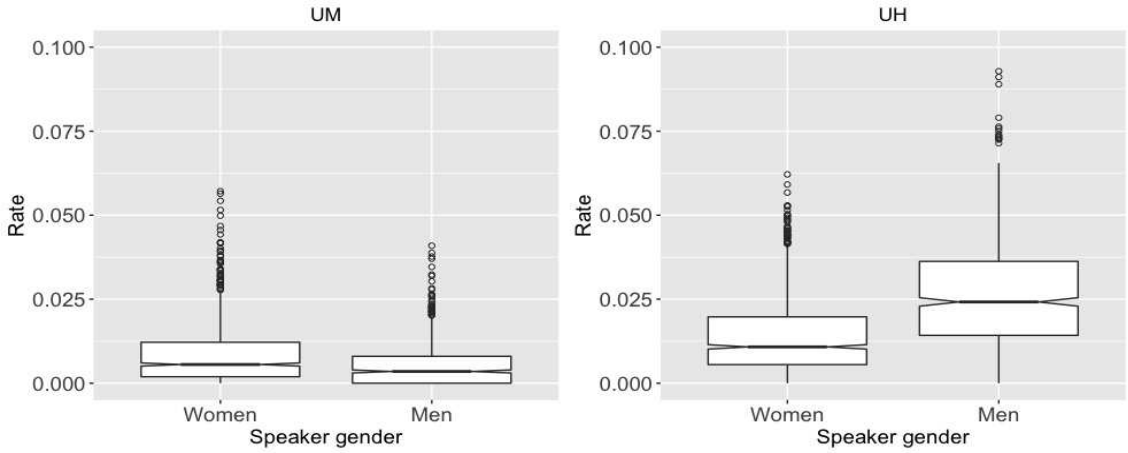

FIGURE 7.17: Effect of speaker gender on the production rate of *um* and *uh* - empty circles represent outliers

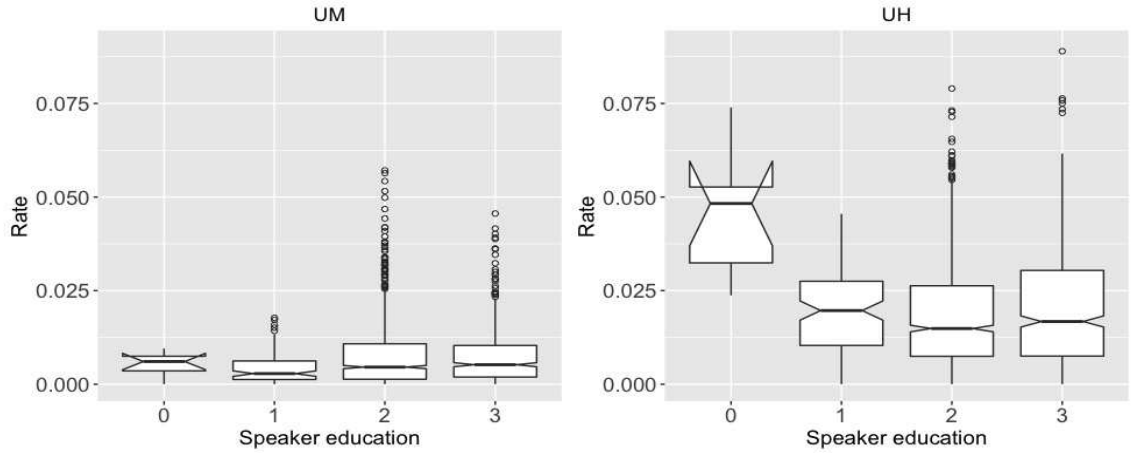

FIGURE 7.18: Effect of speaker education on the production rate of *um* and *uh* - empty circles represent outliers

7.6.4 Summary and discussion

Results from this section indicate that all variables related to the speaker (i.e., speaker, gender, age, dialect, and education level) have a significant effect on the rate of *uh*, and that all variables but education level have a significant effect on the rate of *um*.

The production of *um* vs. *uh* varies with regards to gender, age, dialect, and education level. Different dialects have various effects on the rate of *um* vs. *uh*, education does not reliably affect the rate of *um* but it has a decreasing effect on the rate of *uh* for education levels 0 through 2, and gender and age have an opposite effect on the rates of the two markers. Older speakers use less *ums* but more *uhs* than younger speakers, and women use more *ums* while men use more *uhs*.

These results show that the speaker variable and variables related to the speaker are important to take into account when looking at the rates of *um* and *uh* because they affect the production of the two markers, and they affect them in different ways. This further shows that *um* and *uh* are different variables.

7.7 Position

The goal of this last section is to find out whether the position of *um* and *uh* affects whether transcribers are more likely to miss or hallucinate the marker. The position of the two markers is determined by their position in the slash unit, used to segment speech in the Switchboard transcriptions. The four positions are : alone, initial, medial, or final. Figure 7.19 shows the distribution of *um* and *uh* for each position in the slash unit, for markers that do not have transcription errors. The proportions are computed over the total number of *ums* (6,021) and *uhs* (22,178) that do not have any transcription error, determined by the transcription alignments. Note that the proportions are computed out of raw counts in each position. The proportions show that most markers are in medial position, 53.1% of *ums* and 73% of *uhs*, and very few markers are in isolated position (i.e., alone), 6.1% for *um* vs. 1.2% for *uh*[1]. Proportionally speaking, more *ums* are in initial, final, and alone positions than *uh*, but the trends are similar. These results are not compared to other studies on the position of *um* and *uh* due to the use of different datasets and different methods (i.e., counts vs. rates).

1. Note that these results are different from the position results in the ATAROS corpus discussed in the next chapter. The results from next chapter are computed over a different corpus and are in line with other studies that show most disfluencies happen in initial position or alone, due to higher cognitive and planning loads, floor holding, and turn-taking. Furthermore, given the fact that the Treebank3 transcriptions were made without audio input, it is likely that the slash units are not representative of spoken speech units, and therefore raise questions on whether they should be used to determine position. Further work will explore the use of prosodic units as a more universal way to determine position.

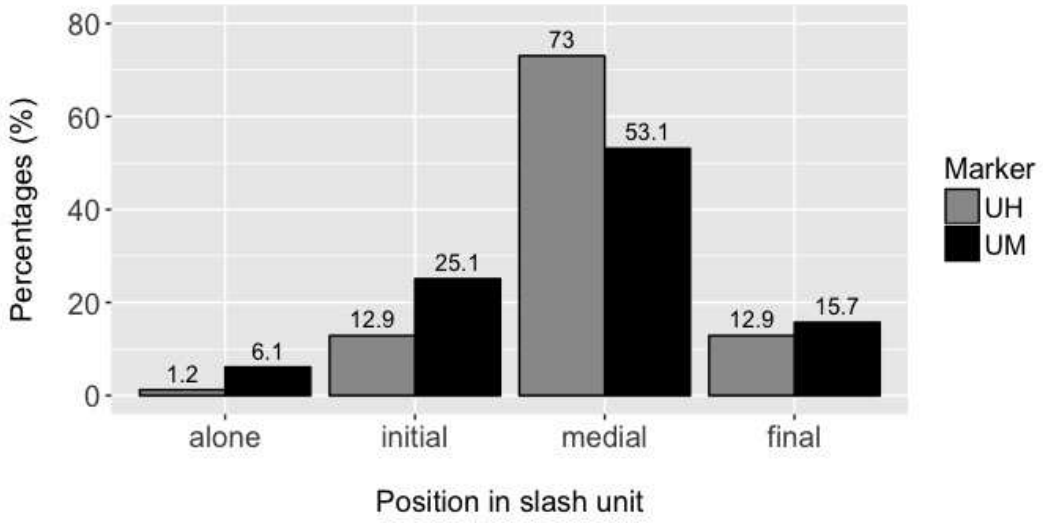

FIGURE 7.19: Proportions of *um* and *uh* with no transcription errors, by position in the slash unit

Figure 7.20 plots the proportions of transcription errors for each marker, relative to the number of markers in each position. The proportions are based on the total number of markers across all transcription errors in the aligned data (9,166 *ums* and 25,950 *uhs*). Results show that markers in all positions are substituted and hallucinated, but only markers in initial and medial position are missed. Initial markers are proportionally the most likely to be missed, approximately twice as much as markers in medial position. Isolated *ums* and *uhs* are proportionally more substituted and hallucinated than markers in other positions, especially for *uh*. The chi square tests for both markers also suggest that position and error type are not independent ($p < 0.001$). These results show that the marker's position and transcription error types are related variables and that transcribers are more likely to make certain types of transcription errors depending on the position of the marker. Finally, it is also interesting to note that position has a similar effect on the error types of both markers, although to slightly different extents.

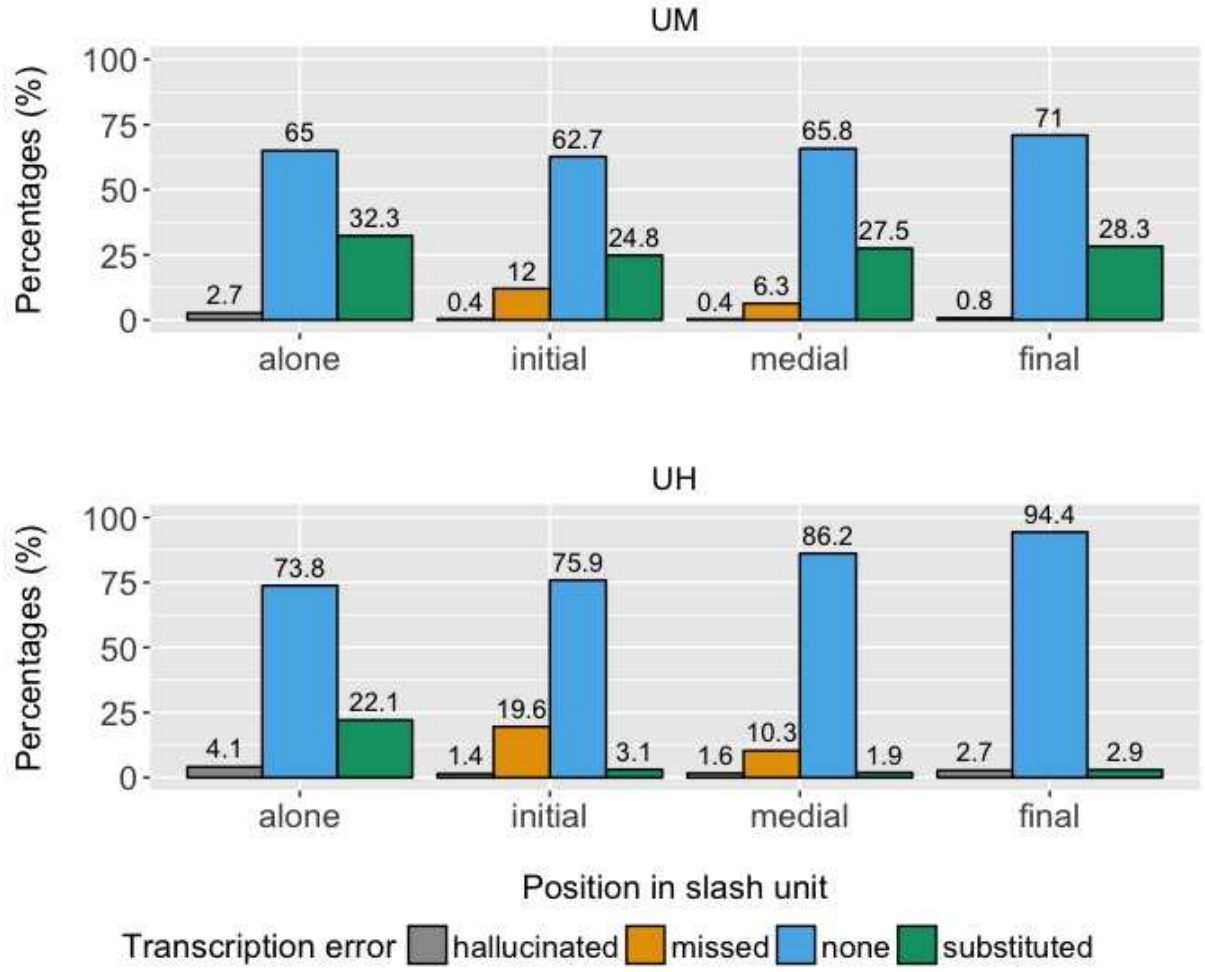

FIGURE 7.20: Proportions of transcription errors for *um* and *uh* for each position in the slash unit, relative to the total number of markers in each position

7.8 Conclusion

This chapter focuses on two main aspects of *um* and *uh*, their production and their transcription errors, both measured in rates to compensate for variability. All sections from this analysis show that *um* and *uh* are affected by various factors, that they generally behave differently from each other, and from other words.

Missed and hallucinated *uhs* represent a larger proportion of all missed and hallucinated

words, and a larger proportion of *uhs*, while missed and hallucinated *ums* make for a smaller percentage of all missed and hallucinated words, and a smaller percentage of *ums*.

The markers *um* and *uh* have different distributions. First of all, there are 2.8 more *uhs* in the corpus than *ums* (25,495 vs. 9,113) and there is more speaker variability for the marker *uh* than *um*. *Um* is substituted more than *uh*, (2,494 vs. 641), but 2,315 substitutions of *um* are by the marker *uh*. After subtracting this substitution type, there are only 179 substitutions of *um* remaining in the 932 conversations, against 641 substitutions for *uh*, substituted 3.6 more times than *um*. In terms of raw counts, *uh* is the most likely word to be missed, and the second most often hallucinated word, against *um* which is the 14th most likely word to the missed, and the 26th most likely word to be hallucinated. Furthermore, proportionally to their token frequency, *uh* is more likely to be missed or hallucinated than *um*. The difficulty and naturalness ratings also affect *um* and *uh* differently. Both difficulty and naturalness have a significant effect on the rate of substituted *ums*, and only naturalness has a significant effect on the rate of missed *uhs*. Transcribers substitute *um* by other words more often in conversations rated more difficult to transcribe, and they substitute more *ums* and miss less *uhs* in conversations that sound less natural. Finally, all variables related to the speaker, gender, age, dialect, and education level have a different effect on the production of the two markers. Education does not have an effect on the rate of *um* but it has a decreasing effect on the rate of *uh* i.e., more educated speakers use less *uhs*. Women use more *ums* than men, and conversely men use more *uhs*. Exclusively different dialects have a significant effect on the rates of *um* and *uh*, and age has an inverse effect on the production of the two markers ; as speaker age increases, speaker use less *ums* and more *uhs*.

The only variable which has the same effect on all words, *um*, and *uh*, is the number of transcriptions made by the transcribers ; as the number of conversation transcribed by the same transcriber increases, the rate of transcription errors decreases.

Results from this analysis also show that there are 4 times more missed words than hallucinated words (27,159 vs. 6,657). Function words tend to be hallucinated more than missed (63.3 vs. 50.5%), lexical words are about as likely to be missed as hallucinated (13.5%

vs. 13.7%), fragments tend to be missed more than hallucinated (14.5% vs. 6.9%), and words from the category *other* also tend to be missed more than hallucinated (21.5 vs. 16.1%). Results taking the percent total of each type into account show that function words are more missed and hallucinated than lexical words, and *other* words are as hallucinated and more missed than function words.

Finally, results on the relation between the marker's position and the transcription error types show that the two variables are related and that depending on its position, a marker is more or less likely to be missed, hallucinated, or substituted. Markers in *alone* and *final* positions are never missed, initial markers are twice as likely to be missed than medial markers, and markers in *alone* position are more likely to be hallucinated or substituted than markers in other positions. These results suggest that depending on the position of the marker, *um* and *uh* have a different salience (i.e., missed markers), or are more likely to be expected by listeners (i.e., hallucinations).

The fact that *um* and *uh* have different distributions and are affected differently, both in terms of production and transcription errors, indicates that the two markers are different. The fact that *um* has less transcription errors than *uh* further suggests that *um* is different from *uh*. Particularly, the fact that *um* is missed less than *uh* suggests that *um* has a higher information load and that it might play a more important role in discourse than its counter part.

The following experiments look at how the discourse and acoustic characteristics of the marker correlate with the attitude of the speaker, to find out whether they vary with speaker attitude and if they behave in similar ways.

Chapter 8

PRESENCE AND POSITION OF *UM* AND *UH* AND STANCE MARKING

8.1 Introduction

The goal of this chapter is to investigate whether the presence and the position of *um* and *uh* in a speech unit affects the probability of the stance marking of the speech unit. The speech unit is called *spurt*, and corresponds to utterances between pauses greater than 500ms (see section 5.3.5 in Chapter 5), and the term *stance* in this dissertation refers to "the speakers' subjective attitudes toward something" (Haddington, 2004, p. 101) (see section 4.1 in Chapter 4 for more background on *stance*). In this experiment I only use the ATAROS corpus because, unlike Switchboard, it is annotated for stance.

As mentioned in section 4.3, *um* figures among the 25 word unigrams selected by the classifier for stance strength recognition (Levow et al., 2014), which suggests that *um* is a key lexical feature in assigning stance strength labels to spurts. This means that the presence of *um* in a spurt is therefore relevant to its stance marking. In addition, the position of the marker is relevant to several dimensions (e.g., salience, likelihood of disfluency, function, etc.). Based on this and prior studies that suggest that the markers *um* and *uh* have key functions in discourse, I hypothesize there is a correlation between the presence and the position of the two markers in a spurt and the stance marking of the spurt. Furthermore, based on results from Chapter 6 and studies described in section 3.1.6, I argue that *um* and *uh* are separate factors and affect differently the probabilities of the stance marking of the spurt.

8.2 Methodology

8.2.1 Data

The data for this experiment come from the ATAROS corpus, detailed in section 5.3 of Chapter 5. The ATAROS corpus is chosen for this experiment because it was designed to elicit various degrees of stance-taking acts in collaborative tasks, with high quality recordings. The data initially consists of 10,405 spurts that contain a total of 1,032 markers : 593 *ums* and 439 *uhs*. Note that for this experiment all markers in spurts labeled with uncertainty or pronunciation issues are excluded from the data (see examples (1) (2), and (3)), which is why there are 33 fewer tokens than in the prior experiment.

(1) 'Up {PRN uh} to you!

(2) Okay. Yeah.. NVC .. (Uh)

(3) Probably next to toilet paper, (um) -

The majority of spurts (90.78%) do not contain any marker *um* or *uh* (9,445). There are 1032 markers in total, and 960 spurts (9.23%) contain at least one marker. Out of the 960 spurts that contain at least one marker, 59 spurts (0.57%) contain more than one marker, and 901 spurts (8.6%) contain exactly one marker. Out of 59 spurts that contain more than one marker, 48 spurts (0.46%) contain two markers, nine spurts (0.09%) contain three markers, and two spurts (0.02%) contain four markers, which is the maximum number of markers per spurt. The percentages are computed over the total number of spurts (10,405). These numbers show that very few spurts contain more than one marker (0.57%), and even fewer contain more than two markers (0.11%). Since the proportions of spurts with more than one marker is very small I do not take into account the number of markers in the spurt in this experiment. Instead, I focus on the presence or the absence of markers and therefore only look at spurts that have only one marker, either *um* or *uh*.

The data for this experiment therefore consist of 10,346 spurts, 9,445 spurts with no

markers (91.29%), 901 spurts that contain exactly one marker (8.71%), of which 374 spurts have one *uh* (3.62%), and 527 spurts have one *um* (5.09%) (see Table 8.1). The percentages are calculated over the total number of spurts in this data set, which now consists of 10,346 spurts.

TABLE 8.1: Percentages and number of spurts that contain no marker, exactly one *uh* and no *um*, or exactly one *um* and no *uh*

marker	number of spurts	percentages
no marker	9,445	91.29%
um	527	5.09%
uh	374	3.62%

It is also interesting to note that the proportions of *um* and *uh* are equal in spurts that contain more than one marker (36 *ums* and 36 *uhs*), and that there is no marker *um* in the two spurts that have four markers, as illustrated in examples (4) and (5). It not possible to draw any conclusions on spurts that contain multiple markers due to the small sample of observations.

(4) *Uh* k *uh* kim *uh* what is it credit union *Uh* you know thats all federal regulation so if theyre at all regulated by the Feds in any way shape or form

(5) Well right now weve got *uh* let me see theres a *uh* theres a *uh uh* a bug that is going around removing all deciduous trees in the United States

8.2.2 Stance strength and polarity

Stance is measured in two dimensions, stance strength and stance polarity, and is annotated at the spurt level by trained annotators (see section 5.3.4 in Chapter 5 for more

information on stance annotations and inter-annotator agreement). Stance strength corresponds to the level of stance in the spurt and can be measured in two ways; binary stance (stance (1) vs. no stance (0)) or in four folds (no stance (0), weak (1), moderate (2), and strong (3) stance). Stance polarity is measured in three ways, neutral (0), negative (-) or positive (+). The notation "x" denotes that annotators were not able to determine the stance values of the spurt. Figure 8.1 summarizes the proportions of spurts with stance strength and polarity markings. The proportions are computed for each stance strength and polarity notation over the total number of spurts (10,346). Note that the proportions differ from Levow et al. (2014) because undetermined stance markings are included here to see the proportions, and spurts with multiple markers are excluded. The proportions show that almost half of the spurts are marked with weak stance, about a fourth of the spurts are marked with moderate or no stance, and less than 1% of the spurts are marked with strong stance. The majority of spurts are marked with neutral stance, more than a fourth of the spurts are marked with positive stance, and 4.77% of spurts are marked with negative stance. Finally, 5.79% of spurts are marked as undetermined. The results show that weak and neutral stance are the most common stance notations over the corpus, while strong and negative stance are the least common, even compared to undetermined stance markings. These results also indicate that spurts with negative and strong stance might need to be omitted in certain analyses due to the insufficient number of samples. In the case of stance strength, strong stance is included in the binary categorization of stance (absence vs. presence of stance).

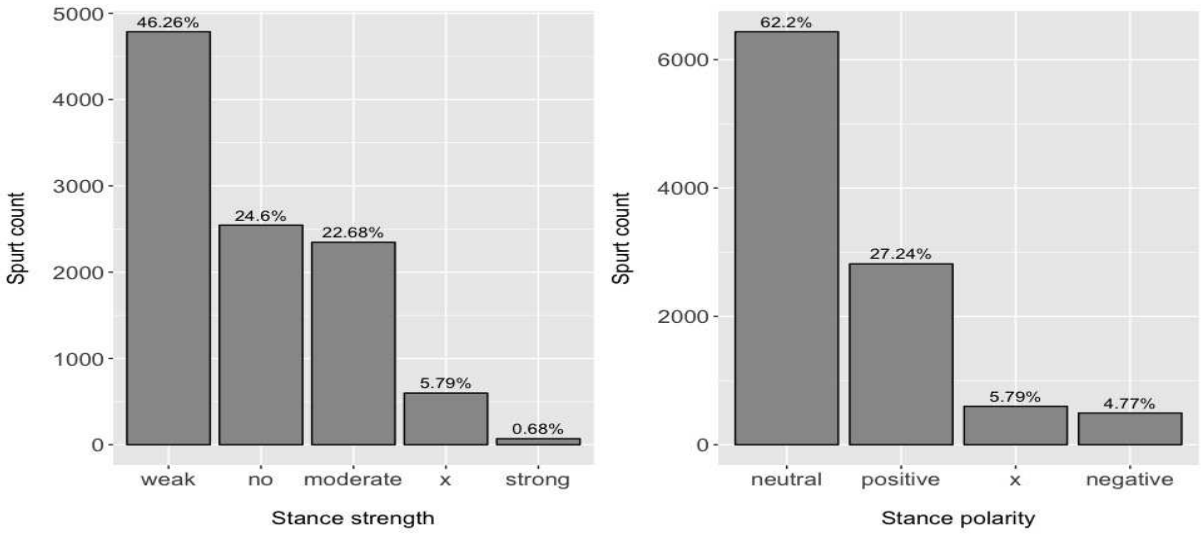

FIGURE 8.1: Counts and proportions of spurts for stance strength (left) and stance polarity (right) - proportions are marked on top of the bars while counts are on the y-axis

8.3 Presence of um and uh

The goal of this section is to find out whether spurt types (i.e., spurts with no markers called *none*, spurts with one *uh*, and spurts with one *um*) are more likely to be marked with a certain stance strength or stance polarity. We know from Figure 8.1 that the majority of spurts are marked with weak neutral stance. In order to compensate for the different proportions of stance markings, Figure 8.2 plots the proportions of each spurt type, computed for its own total (see Table 8.1).

8.3.1 Stance strength

The left side of Figure 8.2 plots the proportions of spurts for each spurt type and for each level of stance strength. The proportions show that spurts with no markers (*none*) are the most likely to be marked with weak stance (47.9%) while spurts with one *um* are the least likely to be marked with weak stance (27.7%). Spurts with one *um* are the most likely to be marked with no stance (48.2%) while spurts with no markers are the least likely to be

marked with no stance (23%). Spurts with one *uh* are more likely to be marked with moderate stance (29.1%) than spurts with one *um* (19.5%), or spurts with no markers (22.6%). The proportions of spurts marked with undetermined stance and strong stance are similar, and vary of less than 2% for undetermined stance and of 1% for strong stance. Results were submitted to a chi square test and show that spurt type and stance strength are dependent or have some association ($p < 0.01$) and provide evidence to suggest that the spurt types behave differently in terms of stance strength.

8.3.2 Stance polarity

The right side figure (Figure 8.2) plots the proportions of spurt types for each polarity level. Proportionally, spurts with one *um* are the most likely to be marked with neutral stance, and spurts of the type *none* are the least likely to be marked with neutral stance. Spurts with one *uh* have the highest proportions of negative labels while spurts with one *um* have the lowest proportions of negative labels. Spurts of the type *none* are the most likely to be marked with positive stance than spurts with one *um* or *uh*, with *uh* being the less likely. Finally, the proportions of spurts marked with undetermined stance are similar and vary by about 1% for each group. The chi square test shows spurt type and stance polarity are not independent ($p < 0.01$), which suggests spurt types behave differently with regards to stance polarity.

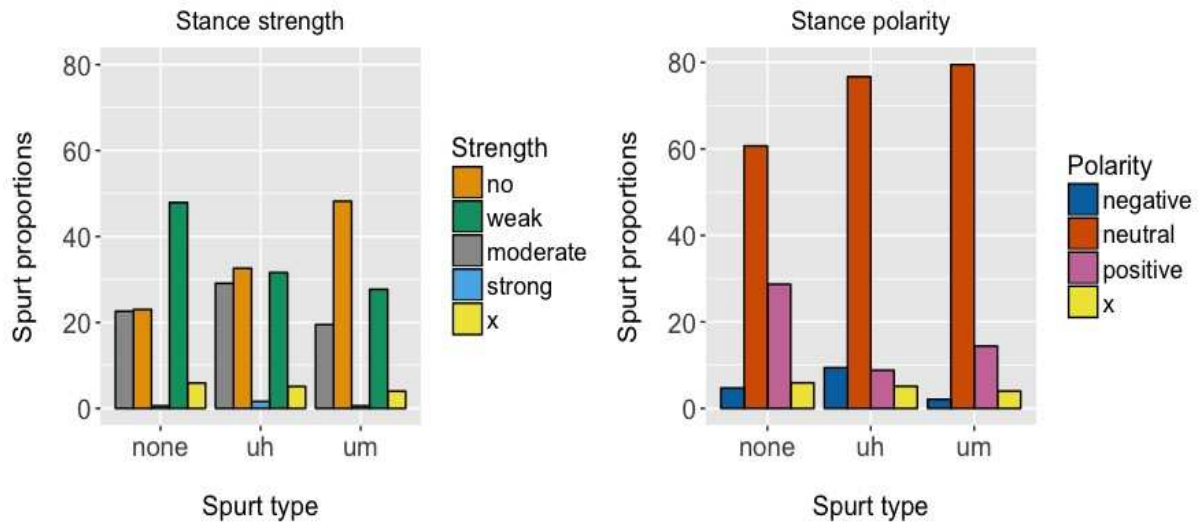

FIGURE 8.2: Proportions of spurts for stance strength (left) and stance polarity (right) for each spurt type (no marker, one *uh* and one *um*)

8.4 Position

The goal of this section is to figure out whether the position of the marker correlates with the stance marking of the spurt in which the marker is. The position variable is relative to the spurt and has four values : alone, initial, medial, and final. The position *alone* means the spurts consists of the marker, and is surrounded by pauses greater than 500ms. Figure 8.3 plots the proportions of *um* and *uh* for each position relative to the spurt, and shows that the two markers have different distributions. The majority of *uhs* (35%) are in medial position and the minority of *ums* (20%) are in the same position. The majority of *ums* are their own spurt (alone) (30%), against 18% for *uhs*. These results show that *um* and *uh* have different distributions in terms of position in the spurt. We saw in the previous section that the presence of *um* and *uh* in a spurt is not independent of the stance marking of the spurt. The next two sections investigate whether their position is dependent on the spurt marking, and to what extent.

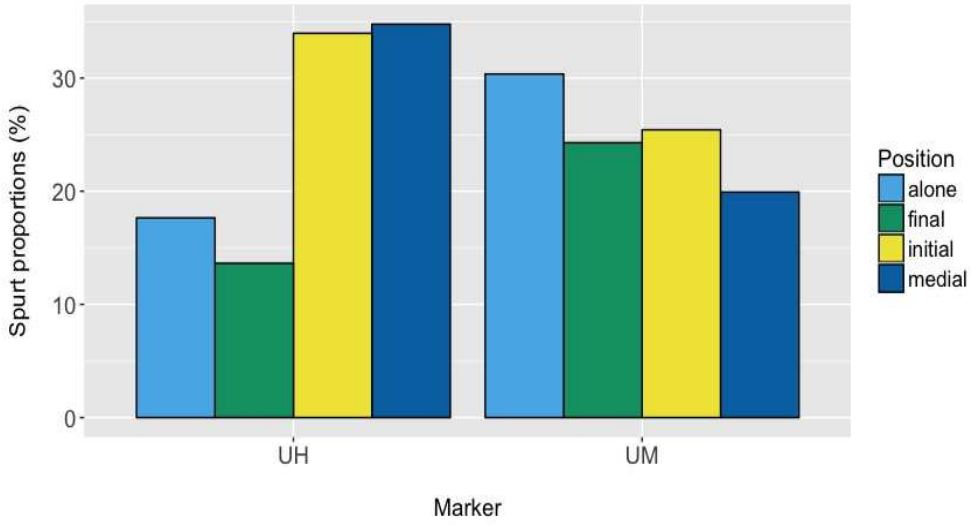

FIGURE 8.3: Percentages of each marker (*um* vs. *uh*) in each position relative to the spurt (alone, initial, medial, and final)

Since there are very few spurts labeled with undetermined stance, and since the presence of *um* or *uh* does not seem to affect this label, I exclude spurts labeled as undetermined in this section. The number of spurts after excluding undertimed stance is now 506 for *um* and 355 *uh*.

8.4.1 Stance strength

Since there are four stance strength values (no stance, weak, moderate, and strong stance), but very few spurts labeled with strong stance, and since there are four positions (alone, initial, medial, and final) for a relatively small number of markers, I only look at a binary distinction of stance (no stance vs. any degree of stance) in this section. Figure 8.4 plots the percentages of *ums* (left) and *uhs* (right) by position in the spurt and by binary stance (no stance (0) vs. any degree of stance (1)). The percentages are computed over the total of spurts for each marker (506 for *um* and 355 for *uh*). The results show that isolated (alone) *ums* and *uhs* are primarily marked with no stance (29% for no stance vs. 1% for any stance

for *um*, and 15% vs. 3% for *uh*). However, spurts that have one *um* or one *uh* in other positions are more likely to be marked with any degree of stance than no stance, with the biggest difference for spurts with markers in medial position (16% with any stance vs. 4% for no stance for *um*, and 31% vs. 4% for *uh*). The positions with the least difference in terms of proportions are initial *ums* (15% for any stance vs. 11% for no stance), and final *uhs* (10% for any stance vs. 3% for no stance). The chi square tests for the two markers corroborate the positions of the two markers and the stance strength labels are not independent variables ($p < 0.01$).

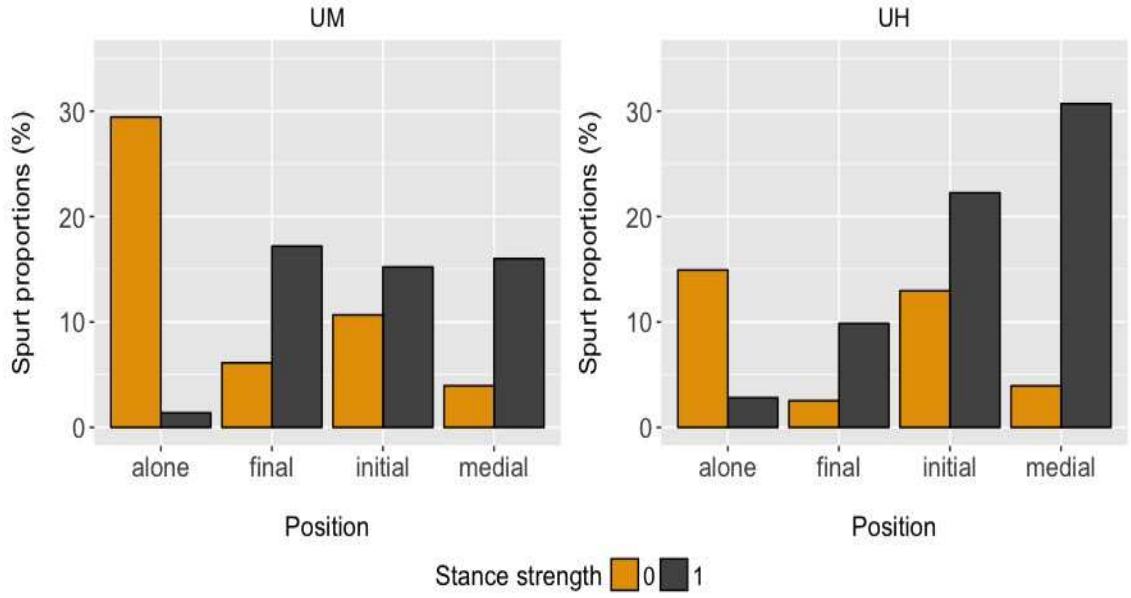

FIGURE 8.4: Percentages of spurts that contain the marker *um* or *uh* for each position (alone, final, initial, and medial) and a binary distinction of stance strength (no stance (0) vs. any stance (1))

8.4.2 Stance polarity

Results plotted in Figure 8.5 show that no matter the position of the marker, most spurts that contain a marker are labeled as *neutral*. These results are not surprising since the vast

majority of spurts that contain an *um* or an *uh* are labeled with neutral stance polarity (see section 8.3.2). Results on the proportions of stance polarity labels based on the markers' position show that isolated *ums* (alone) are all marked as neutral, and that spurts with final and medial *ums* have the highest percentages of positive stance labels (6.5%). Results also show that spurts with initial *uhs* have the highest proportions of neutral stance (32%), and spurts with final *uhs* have the highest proportions of negative (5%) and positive stance (5%). Results were submitted to a chi square test which indicate that the markers' position in the spurt and the stance polarity labels of the spurts are dependent variables ($p < 0.01$).

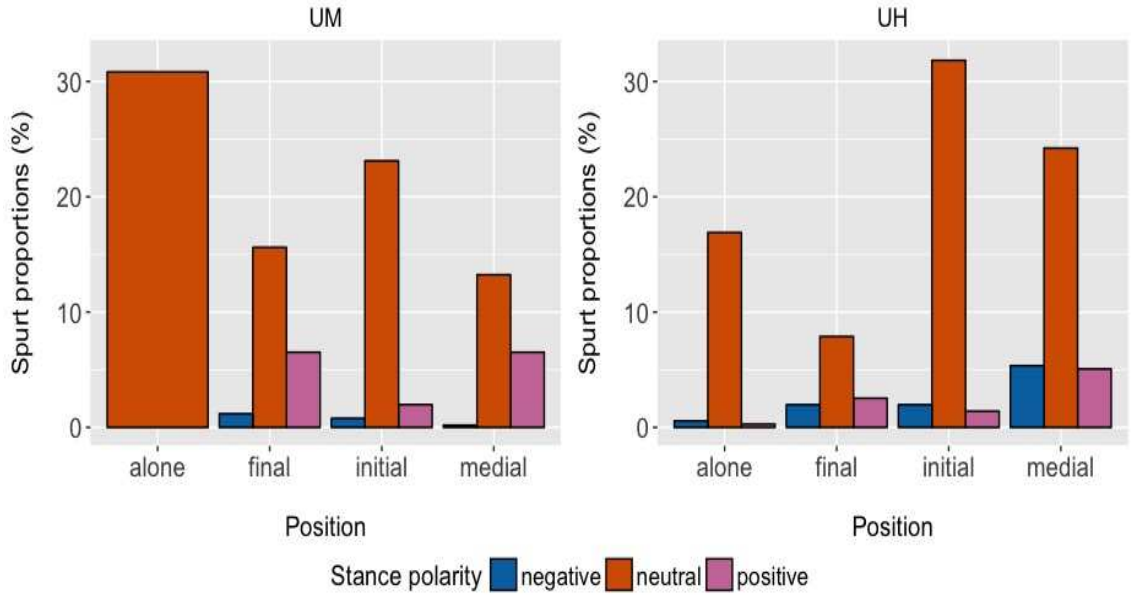

FIGURE 8.5: Percentages of spurts that contain the marker *um* or *uh* for each position (alone, final, initial, and medial) and each stance polarity value (negative, neutral, positive)

8.4.3 Spurts with more than one marker

The goal of this section is to explore how the presence of multiple markers and their position in the spurt is related to the stance marking of the spurt. There are 59 spurts that contain more than one marker (see section 8.2.1), but three are excluded because they are

marked with undetermined stance, which leads to a total of 56 spurts for this section of the experiment. Figures 8.6 and 8.7 summarize the distribution of the stance strength and polarity labels of the spurts that contain more than one marker. Since the sample size for this experiment is small, results are based on the counts of spurts rather than on the proportions. Compared to spurts that contain only one marker, spurts with more than one *um* or *uh* are more likely to be marked with some degree of stance but are proportionally similar in terms of polarity. That is, the majority of spurts that contain more than one marker are more likely to be marked with neutral stance, while spurts with medial *ums* and *uhs* are more likely to be marked with positive polarity, and spurts with medial *uhs* are more likely to be marked with negative polarity.

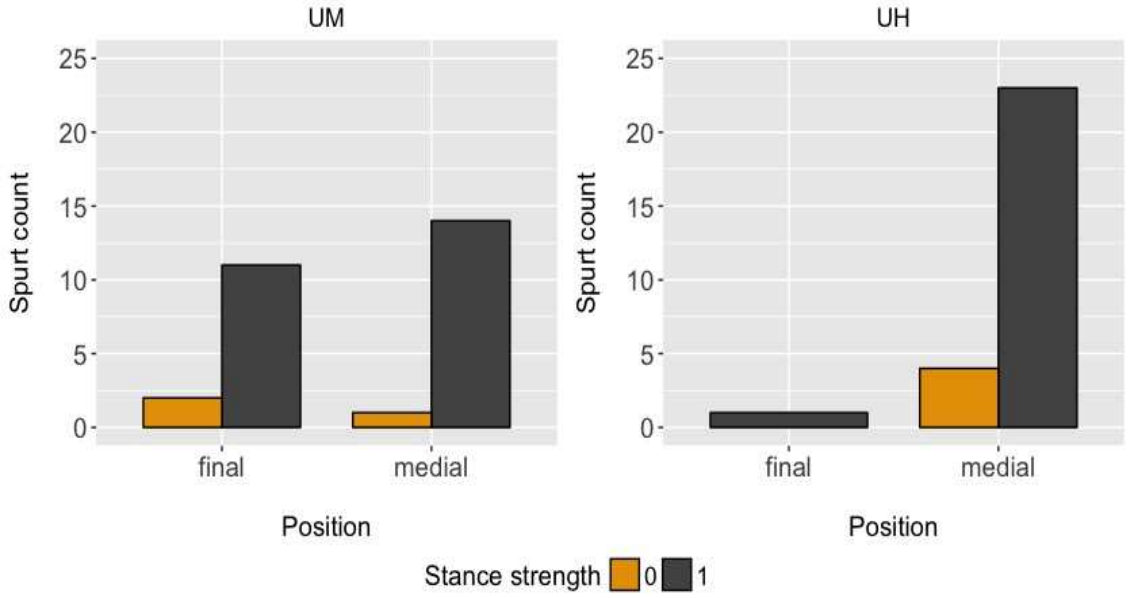

FIGURE 8.6: Percentages of spurts that contain more than one marker for each position (alone, final, initial, and medial) and a binary distinction of stance strength (no stance (0) vs. any stance (1))

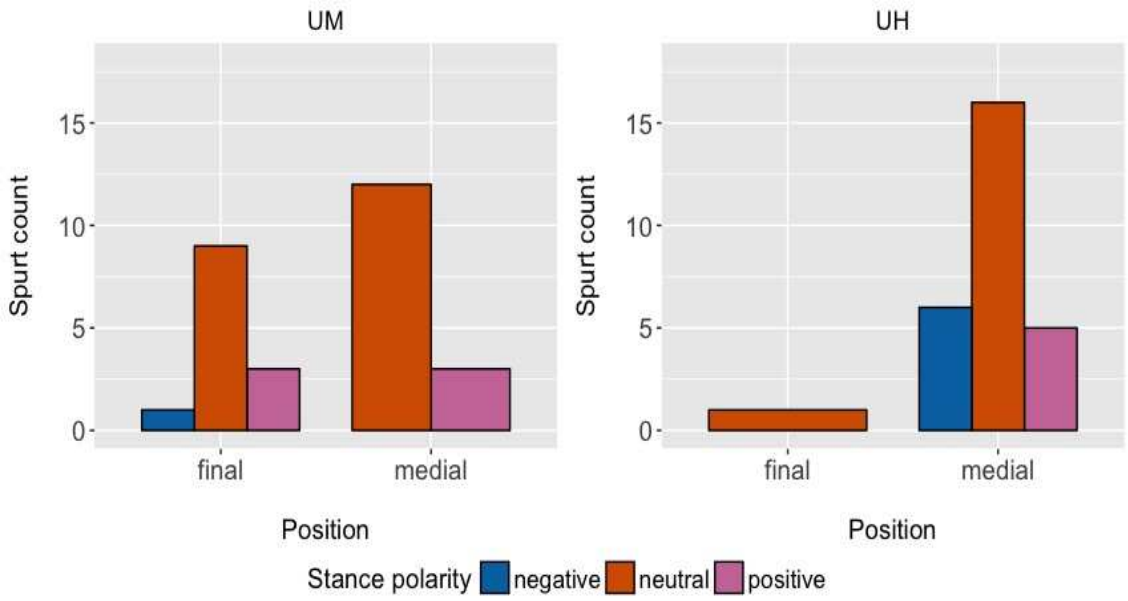

FIGURE 8.7: Percentages of spurts that contain more than one marker for each position (alone, final, initial, and medial) and each stance polarity value (negative, neutral, positive)

8.5 Position in ATAROS vs. Switchboard

In this section I discuss the position of *um* and *uh* in Switchboard and in ATAROS, and the questions it raises about segmenting transcriptions in utterances (i.e., slash units vs. spurts). Results from section 7.7 in Chapter 7 show that the vast majority of *um* and *uh* are in medial position (53.1% and 73% respectively), and very few markers are alone (1.2% and 6.1% respectively). Similarly, results from section 8.4 show that the majority of *uhs* are in medial position in ATAROS, however, with very different proportions, 35% in ATAROS vs. 73% in Switchboard. Other similarities between the position of the two markers in the two datasets are that about 13% of *uhs* are in final position, and about 25% of *ums* are in initial position. Differences between the two datasets are that alone and initial *uhs* in ATAROS are three times more frequent than in Switchboard, alone *ums* are five times more frequent in ATAROS than in Switchboard, and there are more than two times fewer *ums* in medial position in ATAROS than in Switchboard. Another important difference between the two

corpora is that *um* and *uh* have rather similar distributions in Switchboard, unlike ATAROS. For instance, in Switchboard, the majority of *ums* and *uhs* is in medial position, and the minority is alone. On the contrary, in ATAROS, the majority of *ums* are alone while alone is the second to least frequent position for *uh*, and the minority of *ums* are in medial position while the majority of *uhs* are in medial position.

As mentioned in section section 7.7, the Switchboard transcriptions used in this study were segmented into slash units (sentence-like chunks, see section 5.4.3 for more information) from transcriptions only, with no audio input. On the other hand, the ATAROS corpus was segmented based on audio inputs, and spurts are defined as time-units for stance annotation (Freeman, 2015). Since position is measured relative to spurts and slash units which are different types of speech units, it is therefore not possible to compare the position of the markers across corpora. This raises issues on transcription segmentation for speech. Unlike written language that has full sentences, speech does not always have finite, systematic sentences. Many speech sentences can be complete and yet not sentential. Speech is also characterized by interruptions and disfluencies, which lead to partial or incomplete sentences. It would therefore be interesting to determine a universal speech unit, most likely based on the prosodic phrase, to segment speech corpora. This would allow comparing results across datasets, and to increase the robustness of results such as those presented in this study.

8.6 Summary and discussion

To summarize findings from this chapter, spurts with one *um* or *uh* are more likely to be marked with no stance (strength and polarity), and less likely to be marked with weak and positive stance than spurts that do not contain any marker. Spurts with *uh* are the most likely to be labeled with moderate and negative stance. Finally, all spurt types have similar proportions of strong stance and undetermined stance. The results indicate that spurt type and stance strength and polarity are dependent ($p < 0.01$), and spurt types behave differently with regards to stance marking. It is also important to underline that spurts with one marker behave differently from each other, and from spurts with no markers.

Results on the relation between the stance strength label of the spurt and the position of *um* and *uh* indicate that isolated *ums* and *uhs* are more likely to be marked with no stance and neutral polarity, while spurts that contain markers in other positions are more likely to be marked with some degree of stance strength, and negative or positive polarity. The results indicate the marker's position and stance (strength and polarity) are dependent variables ($p < 0.01$). That is, spurts behave differently with regards to stance marking depending on the marker's position. Results on spurts that contain more than one marker suggest that the spurts behave differently in terms of stance strength, but not in terms of stance polarity.

This analysis suggests the presence and the position of the marker in a spurt have an association with the stance label of the spurt. A possible interpretation of these results would suggest that isolated *ums* and *uhs* are less likely to indicate the speaker's attitude than markers in different positions in the spurt, which means that markers in initial, medial, and final position are more likely to play a discourse function than isolated markers. I argue that a more reliable interpretation of the results, however, suggests that the presence and the position of *um* and *uh* can be used as a predictor of the stance label of the spurt.

In the next chapter, I investigate the effect of stance on the acoustic properties of *um* and *uh*, to see if similarly to the presence and the position, there is any correlation between the acoustic realization of the two markers and the stance label of the spurt in which they are. Results from this chapter and from the next chapter are incorporated into a stance classification experiment (see Chapter 10), to test whether the lexical, discourse, and the acoustic features of the two markers improve stance prediction accuracy.

Chapter 9

STANCE VALUES AND ACOUSTIC PROPERTIES OF *UM* AND *UH* IN THE ATAROS CORPUS

9.1 Introduction

The two goals of this experiment are to find out whether *um* and *uh* have different acoustic realizations depending on the stance marking of the utterance in which they are, and whether the two markers behave differently from each other and from other monosyllabic words with the same vowel.

The previous experiment in Chapter 8 shows that the presence and the position of the markers *um* and *uh* in a spurt correlate to some degree with the stance strength and polarity label of the spurt. In addition, results from Levow et al. (2014) show that *um* is a key lexical feature in automatic stance strength recognition. These two experiments suggest that the lexical features *um* and *uh*, and stance labels are dependent variables. Other studies on the acoustic properties of stance marking also suggest that stance is also characterized by acoustic information. Results from Somasundaran et al. (2006) show annotators perform better in annotating opinion categories when they have access to recordings than when they have access to transcriptions only. Experiments from Freeman (2014); Freeman et al. (2015b) show that different functions of *yeah* differ in terms of combination of pitch and intensity contours, as well as vowel duration. Finally, results from Freeman (2015) on stressed vowels in content words show that pitch and intensity increase with stance strength, positive polarity is mainly characterized by longer vowels, while formants do not seem to indicate any particular stance category.

Based on these findings, I investigate whether the stance label of the spurt has an effect on the acoustic realization of the markers *um* and *uh*, and on the same vowel (/ʌ/) in

other monosyllabic words. I hypothesize that *um* and *uh* have different acoustic properties depending on stance markings, and that *um* and *uh* have different acoustic realizations from other monosyllabic words with the same vowel. Based on results from Chapters 6, 7, and 8, I also anticipate that *um* and *uh* vary in different ways depending on stance labels, and that *um* varies to a greater extent than *uh*. Each section of this experiment focuses on one acoustic feature of the vowel /ʌ/ in monosyllabic words. The acoustic parameters are duration (s.), pitch (Hz), intensity (dB), F1 (Hz), and F2 (Hz). For each acoustic feature, I investigate the effect of stance and word group (*um*, *uh*, or *other*). I use three stance measures : binary stance strength, three-way stance strength, and stance polarity (see section 9.2.1 for more information).

9.2 Methodology

9.2.1 Stance data

The stance polarity and stance strength labels are collected from the coarse tier of each TextGrid (see section 5.3.4 in Chapter 5). The stance labels are annotated by trained transcribers, with high inter-rater agreement, 0.87 for stance strength and 0.93 for stance polarity for the weighted Cohen's kappas with equidistant penalties (Freeman, 2015). Stance strength is annotated with four levels : no stance, weak, moderate, and strong stance. In this experiment, I look at two stance strength distinctions, binary stance and three-way stance. The three-way stance distinction includes 3 labels : no stance (0), weak (1), and moderate/strong (2), in order to compensate for the small number of spurts labeled with strong stance. The binary distinction collapses weak, moderate, and strong stance to a category called *any stance*, yielding 2 labels : no stance (0) vs. any degree of stance (1). Stance polarity has 3 levels : neutral, negative, and positive stance. Undetermined stance (x), whether for degree or polarity, is excluded from this experiment due to its small sample size, and in order to reduce the number of stance categories.

9.2.2 Acoustic data

This experiment is conducted on 34 conversations between 17 dyads (groups of 2 speakers). The acoustic data is collected using Praat (Boersma and Weenink, 2015), from intervals in the phone tiers labeled AH1 (/ʌ/ in Arpabet). The phone tiers are created from the utterance tiers with forced alignment using the Penn Phonetics Lab Forced Aligner (P2FA) (Yuan and Liberman, 2008). The vowel boundaries are therefore defined by the forced aligner. The main drawback of this method is that vowel boundaries may have an error margin, which is compensated by using a large number of samples. Acoustic data are automatically collected by a Praat script that runs through the 34 sound files and TextGrids. The script collects acoustic measurements at the midpoint of the vowel for pitch (F0), intensity, F1 and F2, and throughout the interval for vowel duration, based on boundaries from the phone tiers. The script settings are 0.01sec time step, 5 formants, 5,500Hz maximum formant, 0.025sec window length and a minimum pitch of 75Hz. In order to get better measurements, I only used vowels longer than 30ms. Note that throughout this experiment vowels are identified by using the label AH1 from the forced aligner. However, this does not mean that all vowels have the formantic values of AH1. Therefore, the labels *AH1* and /ʌ/ should be considered as broad phonological annotations rather than phonetic markings of the vowels, which allow identifying the vowel boundaries for phonetic measurements, rather than claiming the actual phonetic value of the vowel.

There are 3,902 vowels labeled AH1 from monosyllabic words in spurts not marked with undetermined stance. Note that vowels in words like *check-up* or *uh-huh* are not included because I do not count these words as monosyllabic. Furthermore, 1,615 vowels have undefined pitch (41.5% of all vowels) due to the inability from the pitch extractor to extract the pitch values. The data for pitch therefore consist of 2,273 vowels with 333 *ums* and 198 *uhs*, while the data for intensity, duration, F1 and F2, consist of 3,888 vowels, with 568 *ums* and 416 *uhs*. Table 9.1 summarizes the number of vowels labeled /ʌ/ for each word type, *um*, *uh*, and other monosyllabic words (*other*) for each stance distinction used in this experiment. These

numbers show that there are less vowels in negative and positive spurts, which is expected since there are fewer spurts of these two types in the corpus (see section 8.2.2 in the previous chapter).

TABLE 9.1: Number of vowels labeled /ʌ/ for each stance distinction, three-way and binary stance strength, and stance polarity, and for each word group, *um*, *uh*, and other monosyllabic words (*other*)

stance	value	*other*	*uh*	*um*
3-way strength	no stance	452	128	263
	weak	1078	136	165
	moderate/strong	1374	152	140
binary strength	no stance	452	128	263
	stance	2452	288	305
polarity	negative	328	48	15
	neutral	2140	324	464
	positive	436	44	89

9.2.3 Statistical analysis

The statistical method used in this chapter is Linear Mixed Effect (LME) Model, using the lme4 package (Bates et al., 2015) from R (R Core Team, 2013). Two statistical models are used depending on the research question. I use a random intercept model to look at how the vowel /ʌ/ changes depending on word group (*um, uh* or *other*), with one of the following

acoustic features for dependent variable : duration, intensity, F0, F1, or F2. In this model, the fixed effects are gender and word group, and the random effect is speaker. I use a random slope model to look at how stance affects the vowel /ʌ/ within each word group. Stance is used as the random slope factor, which allows speakers to have different intercepts and slopes for the effect of stance. In other words, this entails that baseline levels of stance could have different effects depending on speakers. The dependent variable is one of the acoustic features (duration, intensity, F0, F1, or F2), and the fixed effects are gender and stance. Stance has three possible distinctions : three-way and binary stance strength, and stance polarity. The random effect is speaker for the word groups *um* and *uh*, and the random effects are speaker and word when looking at *other*. Finally, I use the Likelihood Ratio Test to attain p-values and to test the statistical significance between various models.

9.3 Duration and Stance

This section focuses on testing the effect of stance on the duration of the vowels in three types of words, to see if a) stance strength and polarity affect the duration of the vowel, and b) to find out whether *um* and *uh* behave differently from each other and from other words.

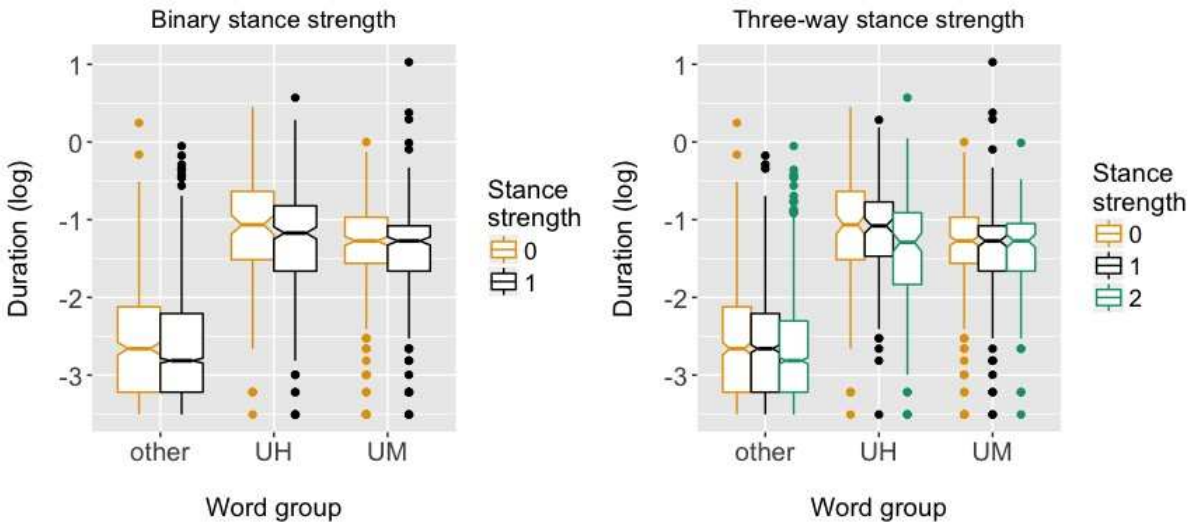

FIGURE 9.1: Effect of stance strength (binary on the left, and three-way on the right) on the duration (in log) for the vowel /ʌ/ in *uh*, *um*, and other monosyllabic words (*other*)

9.3.1 ***Binary stance strength***

Figure 9.1 plots the duration of the vowel /ʌ/ in three word groups (*um, uh* and *other*) depending on binary stance strength (left) and depending on three-way stance strength (right). The left-side figure shows that tokens in spurts marked with some degree of stance have a shorter vowel than tokens in spurts with no stance, and that the difference in duration is stronger for *other* words than for *um* and *uh*.

The statistical analysis of the effect of binary stance strength on vowel duration and the likelihood ratio test between the full model and the reduced ones show that binary stance does not have a significant effect on the vowel of *um*, *uh*, or *other*. However, gender has a significant effect on the vowel of *um* ($\chi^2(1) = 8.2$, $p < 0.01$), increasing duration by 0.321 for men ± 0.096 (standard errors), and on the vowel of *other* ($\chi^2(1) = 10.8$, $p < 0.01$), decreasing duration by -0.140 for men ± 0.035 (standard errors). Gender does not have a significant effect on the vowel of *uh* but it has an increasing effect for men.

These results indicate that even though binary stance does not have a significant effect

on the vowel of the three word groups, vowels are shorter in spurts marked with stance. It is also interesting to note that the duration of the vowel /ʌ/ in *uh* varies more with binary stance than *um*. Finally, it is interesting to see that gender has an increasing effect on *um* and *uh* for men, while it has a decreasing effect for men on other monosyllabic words (*other*).

9.3.2 ***Three-way stance strength***

The right-side figure of Figure 9.1 shows that stance strength has a decreasing effect on the duration of /ʌ/ in the three word groups, especially between categories 1 (weak stance) and 2 (moderate-to-strong stance), and especially for the word groups *other* and *uh*.

The likelihood ratio tests show that three-way stance strength has a significant effect on /ʌ/ in *other* ($\chi^2(1) = 4.5$, $p < 0.05$), decreasing vowel duration by -0.040 $\pm$ 0.017. However, three-way stance strength does not have a significant effect on the duration of /ʌ/ in *um* or *uh*. Results also show that gender has a significant effect for *other* ($\chi^2(1) = 8.4$, $p < 0.01$), decreasing duration for men by about -0.122 $\pm$ 0.037, and for *um* ($\chi^2(1) = 8.8$, $p < 0.01$), increasing duration by about 0.346 $\pm$ 0.095 for men. Gender does not have a significant effect on the duration of /ʌ/ in *uh*, but similarly to *um*, men produce longer vowels.

These results indicate that as stance strength increases, vowel duration decreases, and that the effect of stance is stronger on the vowel in other monosyllabic words than in *um* or *uh*. Similarly to binary stance, men use longer *ums* and *uhs* than women, while they use shorter vowels in other monosyllabic words.

9.3.3 ***Stance polarity***

Figure 9.2 plots the duration of /ʌ/ in the three word groups (*um, uh*, and *other*) depending of the polarity of the spurt in which the vowels are. The plot shows that polarity has a different effect on the vowel depending on the polarity value (positive, neutral, or negative) and depending on the word group. However, negative polarity seems to be associated with shorter duration for all word groups, and positive polarity correlates with higher duration for *um* and *uh*.

The statistical analysis shows that polarity does not have a significant effect on the vowel in any of the three word groups. Gender has a significant effect on the vowel in the *um* group ($\chi^2(1) = 7.7$, $p < 0.01$), increasing vowel duration for men by 0.294 ± 0.096, and on the vowel of other monosyllabic words ($\chi^2(1) = 5.7$, $p < 0.05$), decreasing it by -0.1 ± 0.039. Gender does not have a significant effect for *uh*, but men use longer vowels. Finally, word group has a significant effect on the vowel duration ($\chi^2(1) = 2037.7$, $p < 0.001$), with longer vowels in *um*

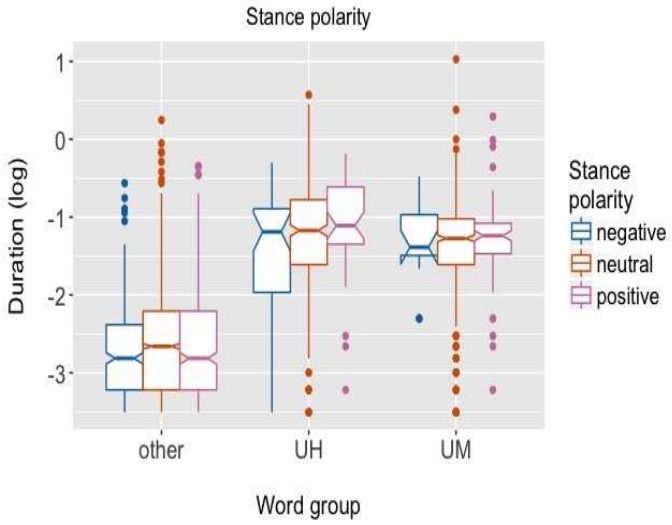

FIGURE 9.2: Effect of stance polarity on the duration (in log) of *uh*, *um*, and other monosyllabic words (*other*)

9.3.4 ***Word group effect***

As illustrated by Figures 9.1 and 9.2, vowel duration varies depending on word group (*other* vs. *um* vs. *uh*). The vowel /ʌ/ is shortest in *other*, and the vowel is shorter in *um* than in *uh*, which is not surprising since *um* has a nasal component. Results also show that the duration of /ʌ/ varies more in *uh* than in *um*, which is surprising since prior results show that *um* varies more than *uh*. The statistical analysis on the effect of word group reports that word group has a significant effect on the duration of the vowel ($\chi^2(1) = 2176$, $p <$

0.001), increasing duration for *um* by about 1.321 ± 0.031, and increasing duration for *uh* by about 1.404 ± 0.035.

9.3.5 ***Summary and discussion***

Table 9.2 summarizes the results on duration. The duration of the vowel is significantly different depending on word group. The vowel is shortest in *other*, and shorter in *um* than *uh*. Gender also has a significant effect on vowel duration. Men use longer vowels for *um* and *uh* than women, while they use shorter vowels than women for other monosyllabic words. Stance has a decreasing effect on vowel duration in the three word groups, and only the three-way stance strength has a significant effect on the vowel of *other*. Based on the statistical analysis and on the medians plotted in Figure 9.1, the three-way distinction is a better measure than the binary distinction, because it captures the differences in duration between weak stance (1) and moderate/strong stance (2), unlike binary stance, which collapses these two distinctions into one category. Unlike for *other*, positive polarity in *um* and *uh* correlates with longer vowels, similarly to findings from (Freeman, 2015). The duration of the vowel in *um* and *uh* varies more than in other words when expressing negative, and especially positive polarity. These results suggest that *um* and *uh* are more susceptible to carry polarity marking than other words.

TABLE 9.2: Summary of the statistical analysis on duration for each word group (*other, uh* and *um*)

factor	*other*	*uh*	*um*
bin	–	–	–
effect	↘	↘	↘
3-way	$p < 0.05$	–	–
effect	↘	↘	↘
polarity	–	–	–
effect	*negative = shortest*		
gender	$p < 0.01$	–	$p < 0.01$
effect (men)	↘	↗	↗
word group	$p < 0.001$		
effect	*other < um < uh*		

9.4 Pitch and Stance

This section looks at the effect of stance and word group on the pitch (F0) of the vowel /ʌ/ in three word groups (*other, uh*, and *um*), for the three stance measures. Figure 9.3 plots the pitch of the vowel depending on stance (binary on the left, three-way on the right) for each word group. The two plots show that the pitch of the vowel varies more in *um* and *uh* than in *other* depending on stance strength.

As expected, gender reliably affects the pitch of the vowel for all three stance distinctions ($p < 0.001$), and men have a shorter pitch than women. Since gender is a well established factor of pitch values, and since it is not the focus of the study, I only report it here.

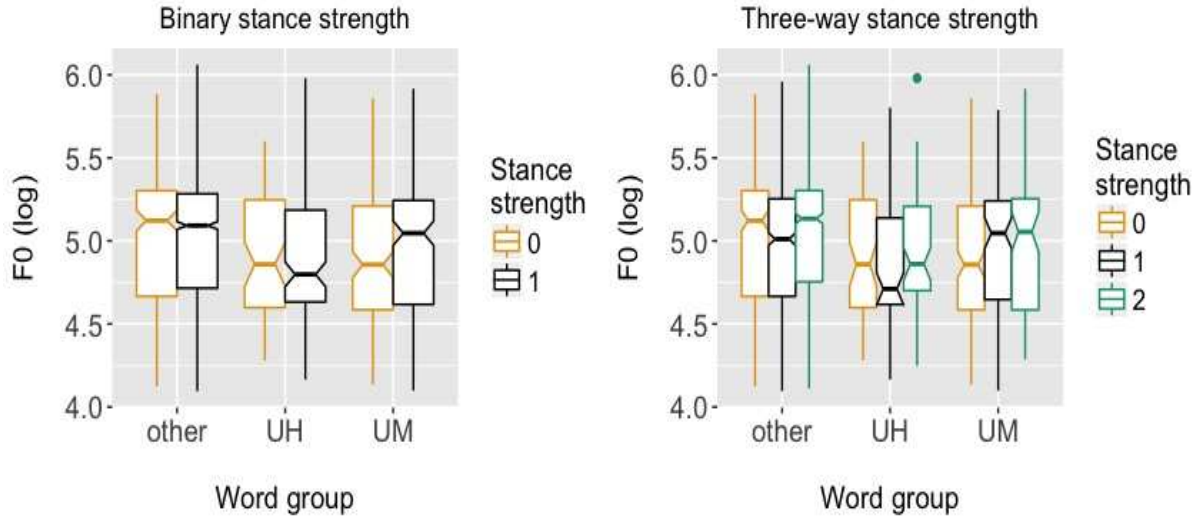

FIGURE 9.3: Effect of stance strength (binary on the left, and three-way on the right) on pitch, also called F0 (in log), for the vowel /ʌ/ in *uh*, *um*, and other monosyllabic words (*other*)

9.4.1 ***Binary stance strength***

Results from the statistical analysis on the effect of pitch on the vowel /ʌ/ show that binary stance does not have a significant effect on the vowel in any word group. The presence of stance (1) has a decreasing effect for vowels in *other* and *uh*, while it has an increasing effect for vowels in *um*.

9.4.2 ***Three-way stance strength***

Similarly to binary stance, the statistical analysis indicates that three-way stance strength does not have a significant effect on the pitch of the vowel in the three word groups. However, the medians plotted in Figure 9.3 show that even though the difference is not significant, the pitch of the vowel varies depending on three-way stance strength. It is interesting to note that for *other* and *uh*, vowels in spurts marked with no stance (0) behave similarly to vowels marked with moderate-to-strong stance (2), whereas vowels marked with weak stance (1)

have a lower pitch. On the contrary, pitch increases for vowels marked with stance in *um*; pitch is similar for weak (1) and moderate-to-strong stance (2), while it is lower for no stance (0).

9.4.3 ***Stance polarity***

Figure 9.4 plots the effect of stance polarity on the pitch of the vowel /ʌ/ in the three word groups. The medians show that the pitch of the vowel in *um* and *uh* varies more than in *other*, especially in *um*. The statistical analysis shows that similarly to stance strength, polarity does not reliably affect pitch. However, the medians for all three groups show that vowels in negative spurts have the highest pitch, and that they tend to have the lowest pitch in neutral spurts. These results suggest that the presence of polarity is marked by a higher pitch than neutral polarity, especially negative polarity.

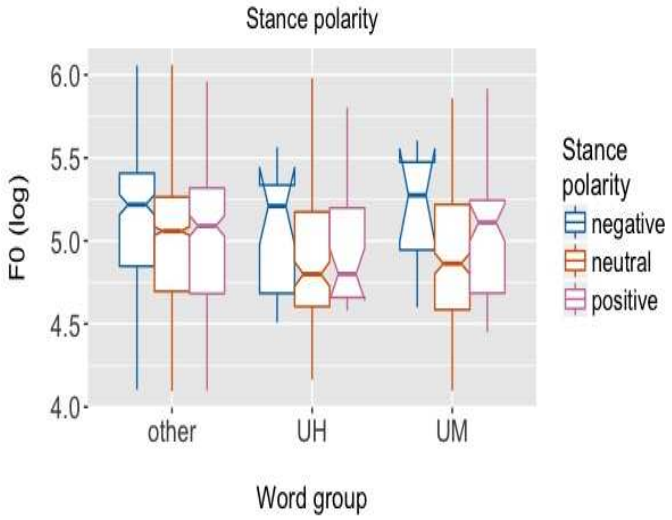

FIGURE 9.4: Effect of stance polarity on the pitch, also called F0 (in log), of *uh*, *um*, and other monosyllabic words (*other*)

9.4.4 ***Word group effect***

The statistical analysis of the effect of word group on pitch shows that pitch is reliably different depending on the word group ($\chi^2(1) = 8.7$, $p < 0.05$). The pitch of vowels in *um* and *uh* is lower than in other by about -0.039 $\pm$ 0.019, and that the pitch of *uh* is on average lower than the pitch of *um*.

9.4.5 ***Summary and discussion***

The results from this section confirm that men have a lower pitch than women, and that the pitch of the vowel is significantly different depending on the word group.

Results for pitch and stance strength indicate that a binary stance distinction is appropriate for *um*, but not for *other* and *uh*. The three-way distinction is better for *other* and *uh* because it captures the differences in pitch between weak and moderate-to-strong stance. The results on the effect of stance strength on pitch corroborate results from Freeman (2015), which show that pitch increases with stance strength, as shown by the medians in Figure 9.3 for moderate-to-strong stance (2). Results on polarity also corroborate findings from Freeman (2015) that find higher pitch on negative *yeah*, and higher pitch as a correlate of positive polarity.

The effect of stance strength and polarity on pitch is different for *um* and *uh*, and *uh* behaves more like *other*. The pitch of vowels in *um* varies to a greater extent, and in different trends across polarity and strength than for vowels in *uh* or *other*. These results suggest that *um* is more likely to signal polarity than *uh* or other words.

TABLE 9.3: Summary of the statistical analysis on pitch for each word group (*other, uh* and *um*)

factor	*other*	*uh*	*um*
bin	–	–	–
effect	↘	↘	↗
3-way	–	–	–
effect	$0 > 1 < 2$	$0 > 1 < 2$	$0 < 1 \approx 2$
polarity	–	–	–
effect	$- > 0 \approx +$	$- > 0 \approx +$	$- > 0 < +$
gender	$p < 0.001$		
effect (men)	men < women		
word group	$p < 0.05$		
effect	*other* > *um* > *uh*		

9.5 Intensity and Stance

In this section I look at the effect of stance and word group on the intensity of the vowel in the three word groups. Figure 9.5 plots the intensity of the vowels in the three word groups depending on binary (left) and three-way (right) stance strength.

Although gender does not reliably affect the intensity of the marker in any of the stance models, results show that men use a slightly higher intensity than female, which varies from 1.6 dB for *um*, to 0.4 dB for *uh*, and to 0.1 dB for *other*. The effect of gender is therefore stronger for *um* than for *uh* and other.

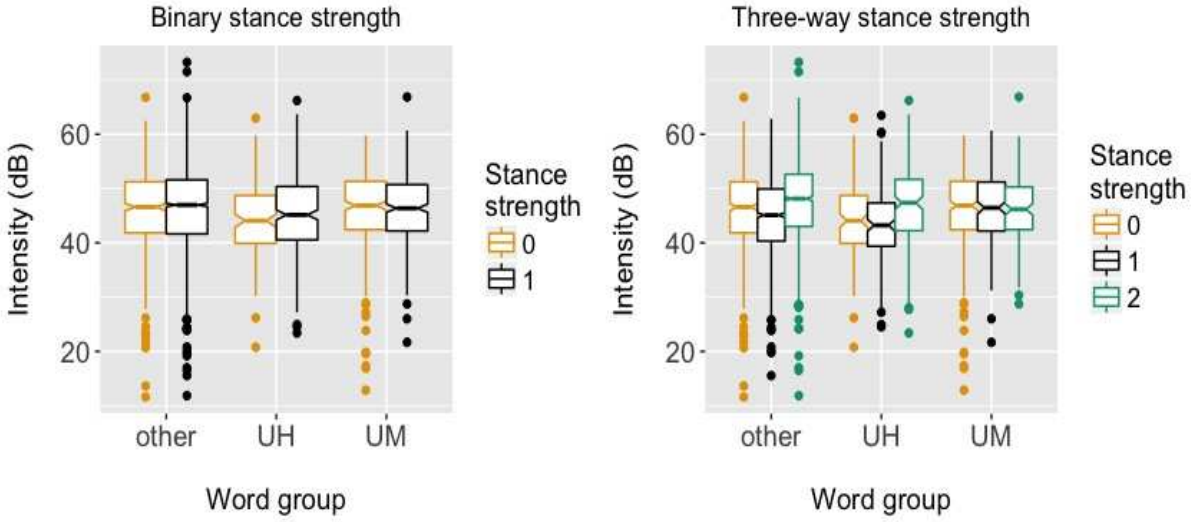

FIGURE 9.5: Effect of stance strength (binary on the left, and three-way on the right) on intensity (in dB), for the vowel /ʌ/ in *uh*, *um*, and other monosyllabic words (*other*)

9.5.1 ***Binary stance strength***

The statistical analysis reveals that binary stance does not have an effect on intensity, and that *um* does not behave like *uh* and *other*. The presence of stance increases intensity for *other* and *uh*, while it decreases intensity for *um*.

9.5.2 ***Three-way stance strength***

Results on the effect of three-way stance strength on the intensity of the vowel report that three-way stance only has a reliable effect on *other* ($\chi^2(1) = 9.3$, $p < 0.01$), with a difference of about 0.9 dB $\pm$ 0.2 between stance values. The medians from the right-side plot in Figure 9.5 show that similarly to pitch, the intensity of the vowel varies with three-way stance in similar trends for *uh* and *other*, while the vowels in *um* behave differently. Similarly to duration, there is very little difference in the intensity of *um* compared to *uh* and *other*. The vowels in *uh* and *other* behave similarly to results reported in Freeman (2015), unlike vowels in *um*, where intensity is rather stable across stance values.

9.5.3 ***Stance polarity***

Results on the effect of polarity on vowel intensity show that polarity does not have a significant effect on intensity in any of the word groups. The plot in Figure 9.6 shows that similarly to stance strength, *um* does not behave like *uh* and *other* with regards to polarity and intensity. For *other* and *uh*, vowels in spurts marked with negative polarity have the highest intensity, while for *um* negative polarity is associated with lower intensity. The results form this analysis suggest that vowels in *other* and *uh* behave more similarly to *yeah* and to other lexical words (Freeman, 2015) than vowels in *um*.

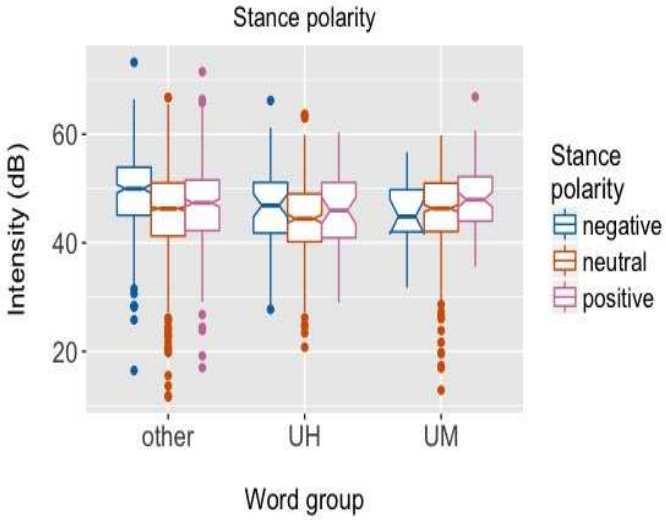

FIGURE 9.6: Effect of stance polarity on intensity (in dB), of *uh*, *um*, and other monosyllabic words (*other*)

9.5.4 ***Word group effect***

The statistical analysis reveals that only vowels in *uh* are significantly different from the other word groups in terms of intensity ($\chi^2(1) = 6.8$, $p < 0.05$), with a difference of about -0.8 dB $\pm$ 0.3.

9.5.5 ***Summary and discussion***

In sum, gender affects the intensity of *um* the most, and word group affects the intensity of *uh* to the greatest extent. Similarly to duration and pitch, *um* does not behave like *uh* and *other* with regards to stance strength, and similarly to duration, the intensity of *um* does not vary much depending on stance strength compared to *uh* and *other*. In addition, polarity affects intensity for *um* in different ways than the two other groups. Results from this section indicate that *uh* and *other* are more similar than *um* when looking at the effect of polarity on vowel intensity. Furthermore, findings also indicate that *uh* and *other* behave more similarly to *yeah* and lexical words (Freeman, 2015) than *um*. This suggests that *um* does not behave like other words, especially lexical words, and that *um* carries different information than *uh*, while *uh* behave more like other lexical words.

TABLE 9.4: Summary of the statistical analysis on intensity for each word group (*other, uh* and *um*)

factor	*other*	*uh*	*um*
bin	–	–	–
effect	$\nearrow$	$\nearrow$	$\approx$
3-way	$p < 0.01$	–	–
effect	$0 > 1 < 2$	$0 > 1 < 2$	$0 > 1 > 2$
polarity	–	–	–
effect	- highest	- highest	- lowest
gender	–	–	–
effect (men)		men < women	
word group	$p < 0.05$	–	–
effect	$\approx$	$\searrow$	$\approx$

9.6 Vowel quality (F1 and F2) and Stance

In this section I look at the effect of stance on the vowel quality in three word groups, and across word group. I measure vowel quality by looking at the first formant (F1) and the second formant (F2) of the vowels. Figures 9.7 and 9.8 respectively plot the first (F1) and second formants (F2) for the three word groups depending on binary (left) and three-way (right) stance strength.

As expected, gender reliably affects F1 and F2 of the vowel in each word group ($p < 0.001$), with a lower value for men than women.

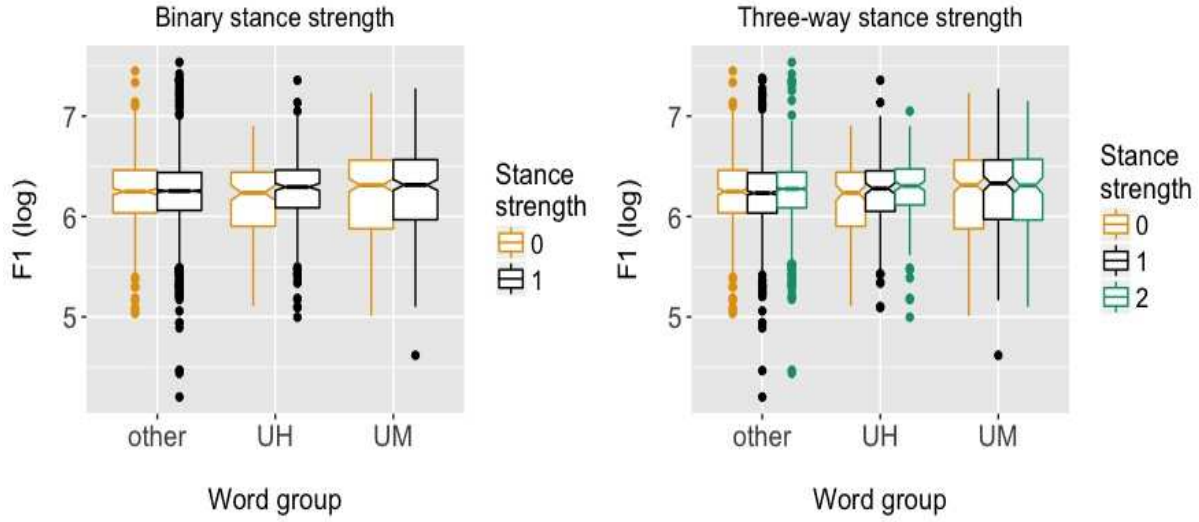

FIGURE 9.7: Effect of stance strength (binary on the left, and three-way on the right) on F1 (in log), for the vowel /ʌ/ in *uh*, *um*, and other monosyllabic words (*other*)

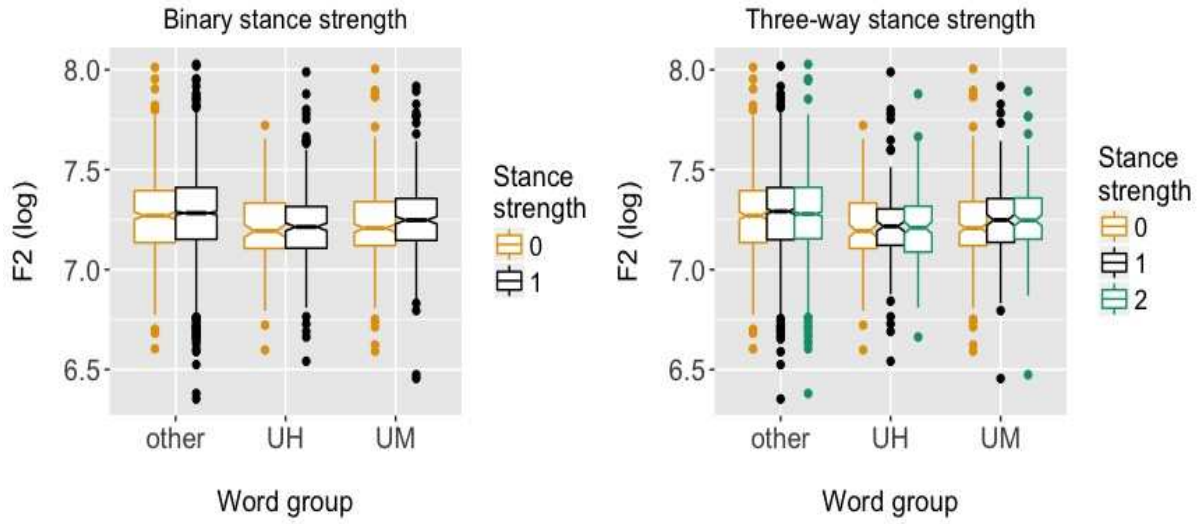

FIGURE 9.8: Effect of stance strength (binary on the left, and three-way on the right) on F2 (in log), for the vowel /ʌ/ in *uh*, *um*, and other monosyllabic words (*other*)

9.6.1 ***Binary and three-way stance strength***

The statistical analysis shows that only three-way stance reliably affects F2 ($\chi^2(1) = 4$, $p < 0.05$) in *other*, and that binary and three-way do not reliably affect F2 in *um* and *uh*, or F1 in any word group. The medians in Figure 9.7 show that compared to other acoustic parameters, F1 does not vary much depending on stance strength, except for *uh*, where F1 increases as stance strength increases. However, F2 increases in the three word groups, and weak stance (1) and moderate-to-strong stance (2) behave similarly. Stance affects F2 in *um* and *uh* in similar ways.

9.6.2 ***Stance polarity***

Results on the effect of polarity on F1 and F2 show that polarity does not reliably affect vowel quality. However, Figure 9.9 indicates that even though the effect is not significant, vowel quality varies depending on polarity, and to a greater extent for F2 than for F1, similarly to stance strength. F1 is higher in spurt marked with negative polarity, especially

for *other* and *uh*, and F2 is higher for negative polarity i *other* and *um*, while it is lower for *uh*. F2 does not behave similarly with regards to polarity for *um* and *uh*, and F2 varies more in *um* than in *uh*.

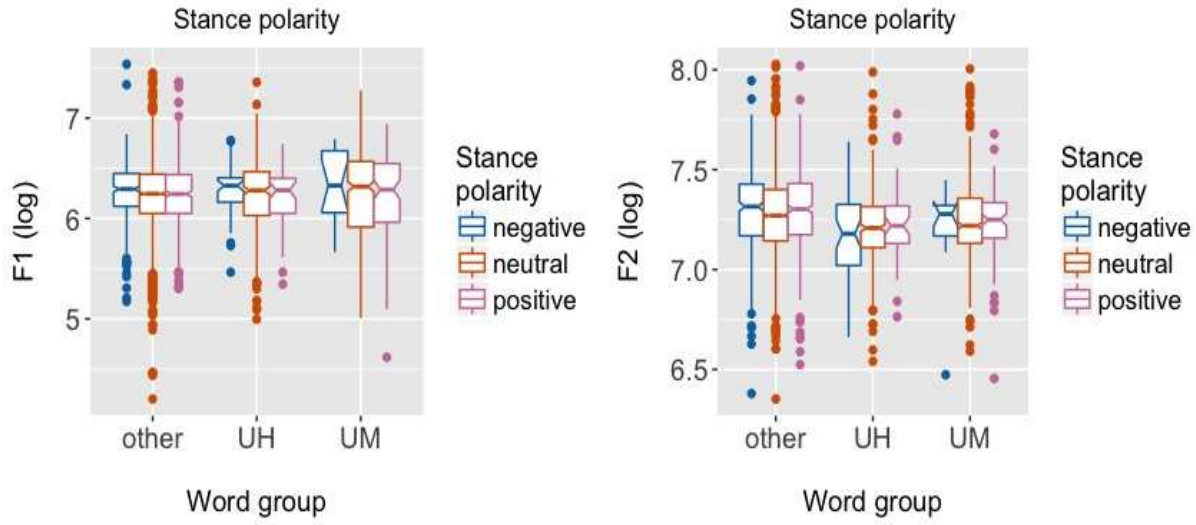

FIGURE 9.9: Effect of stance polarity on F1 (in log) on the left and F2 (in log) on the right, of *uh*, *um*, and other monosyllabic words (*other*)

9.6.3 ***Word group effect***

The statistical analysis shows that F1 is reliably different across word group ($\chi^2(1) = 12.3$, $p < 0.01$). F1 is higher in *uh* by about 0.037 ± 0.017, and it is higher in *um* by about 0.046 ± 0.015. F2 is significantly different depending on word group ($\chi^2(1) = 28.1$, $p < 0.001$), it is lower in *uh* by about -0.02 ± 0.009, and it is lower in *um* by about -0.04 ± 0.008. These results show that vowel quality is different in *um* and *uh* since the vowel is lower and more back in these two word groups.

9.6.4 ***Summary and discussion***

TABLE 9.5: Summary of the statistical analysis on F1 and F2 for each word group (*other, uh* and *um*)

factor	*other*	*uh*	*um*
bin F1	–	–	–
effect	≈	↗	≈
bin F2	–	–	–
effect	↗	↗	↗
3-way F1	–	–	–
effect	≈	↗	≈
3-way F2	$p < 0.05$	–	–
effect	↗	↗	↗
polarity F1	–	–	–
effect		- highest	
polarity F2	–	–	–
effect	- highest	- lowest	- highest
gender F1 F2		$p < 0.001$	
effect (men)		men < women	
word group F1		$p < 0.01$	
effect		*other* < *um* and *uh*	
word group F2		$p < 0.001$	
effect		*other* > *um* and *uh*	

Table 9.5 summarizes the results on stance strength, polarity, gender and word group for vowel quality in the three word groups. Stance strength has an increasing effect for vowels

in *uh*, which means that F1 values increase as stance strength increases, and that the vowel is lower with higher stance strength. Results also indicate that binary stance strength is a good measure for *um* and *uh*, but three-way stance strength captures the difference in F2 between weak and moderate-to-strong stance in the group *other*. Therefore, three-way stance strength is a better measure for vowel quality.

9.7 Chapter summary and discussion

In this experiment I predicted that *um* and *uh* are realized differently depending on the stance of the spurt in which they are. I also predicted that the two markers behave differently from other words, and from each other.

The results presented in this experiment show that stance strength (binary and three-way) and polarity do not reliably affect the acoustic properties of the vowels in *um* and *uh*. However, three-way stance strength reliably affects the duration, the intensity, and the second formant (F2) of other monosyllabic words (*other*). Findings also show that polarity does not have a significant effect on the acoustic parameters of the vowel in any of the three word groups. This is not surprising since other studies have found more correlation between stance strength and lexical or acoustic features, than with stance polarity (Freeman, 2015; Levow et al., 2014).

Even though the results are not significant, interesting findings from this chapter include differences in medians depending on stance strength and polarity for *um* and *uh*, as well as differences in how the two markers are affected by stance for each acoustic parameter. Binary stance has a decreasing effect on the duration of the vowel in all word groups, and three-way stance affects *uh* and *other* to a greater extent than *um*. Binary stance strength affects pitch and intensity differently in the three word groups, it affects F1 to a greater extent for *uh*, and it affects F2 in similar trends, although to a greater extent for *um* than for *uh* and *other*. Three-way stance affects *uh* and *other* in similar ways for pitch, intensity, and F2, and it affects F1 differently in all word groups. Polarity affects duration in different ways, especially for *other*. Polarity affects the pitch and the intensity of *uh* and *other* in similar trends, while

um behaves differently. Polarity does not affect F1 much, and affects F2 in *uh* differently from *um* and *other*. These results show that *um* and *uh* are affected in different ways by the stance marking of the spurt in which they are, and that pitch and intensity follow similar patterns. These results also indicate that *uh* tends to behave more like other monosyllabic words, unlike *um*, which suggests that *um* plays a more important role in stance marking than *uh*, consistent with other findings from this study.

These results suggest that *um* and *uh* are not the same, and that they have different acoustic realizations, although not significant, depending on stance strength and polarity. Findings from this chapter and from Chapter 8 are incorporated in the next chapter to test whether the presence, the position, and the acoustic properties of *um* and *uh* can improve automatic stance classification.

Chapter 10

AUTOMATIC CLASSIFICATION OF STANCE

10.1 Introduction

The goal of this experiment to to test whether incorporating features relating to *um* and *uh* can improve the accuracy of a classifier in predicting stance strength and polarity labels. This chapter takes into account findings from Chapters 6, 8 and 9, and incorporates them together into a concrete applications to see if they can be used in automatic classification of stance.

Results from Chapter 6 show that the speaker gender and the task variables affect the distribution and the duration cues of the two markers, and results from Chapter 9 show that the acoustic characteristics of the vowel in *um* and *uh* have different trends depending on stance strength and polarity labels. Furthermore, results from Chapter 8 show that the presence and the position of the two markers are not independent from the stance label of the spurt in which the marker is, and that the spurts have different probabilities to be label with a certain degree or polarity of stance depending on the presence and the position of the marker. Finally, findings mentioned in Chapter 9 from Freeman (2014, 2015); Levow et al. (2014); Somasundaran et al. (2006) indicate that stance strength and polarity are characterized by different acoustic and lexical characteristics.

Based on these findings, I predict that *um* is a more valuable lexical feature than *uh*, and that incorporating lexical, discourse, metalinguistic, and acoustic features pertaining to *um* and *uh* can increase the accuracy of automatic stance classification over baseline. In the first experiment I test whether *um* and *uh* are relevant word unigram features, and in the second experiment I look at whether metalinguistic, discourse, and acoustic characteristics of *um* and *uh* participate in predicting stance.

10.2 Methodology

The estimator (i.e., the classification algorithm) chosen for this experiment is the SVM package (Support Vector Machine) from Scikit Learn (Pedregosa et al., 2011). This SVM package is a supervised learning method for classification, regression, and outliers detection. It belongs to the discriminant family, and maximizes the margin between two classes. More information on this package is available on the Scikit Learn website : `http://scikit-learn.org/stable/modules/svm.html`. The SVM implementation used for this experiment is SVC (Support Vector Classification), used for the classification of categorical data. The parameters for the implementation of the classifier are the linear kernel, gamma = 0.001, and C = 1.0.

The results are validated by using cross validation, by splitting the data into five folds **X_folds = np.array_split(X, 5)**, where the training data constitutes four folds, the testing data constitutes 1 fold, and the whole process is repeated five times on different training and testing folds (see section D.2 in Appendix D for the code). The accuracy scores are collected in a list for each fold. The estimator (clf) is fitted with the data (X_train) and the target (y_train), to predict the scores based on k folds (5) using the testing set (X_test and y_test), as illustrated in example (1). This methods allows computing the overall mean validity score as well as checking if the training samples are random and if different folds lead to similar predictions. In addition to accuracy scores, prediction, recall, and F1-scores are also used and reported to estimate the best model and the contribution of the features.

(1) scores.append(clf.fit(X_train, y_train).score(X_test, y_test))

The data is extracted via scripts in Praat (Boersma and Weenink, 2015) to collect text information from the spurts, the stance labels, the vowel labels, and the acoustic features. The data for each of the two experiments is organized into dataframes and preprocessed via python scripts. The text preprocessing for the spurts includes stripping punctuation, removing truncated words, removing metalinguistic comments that indicate laughing or other noises, and tokenizing the data. The feature preprocessing steps include exporting all features in the

type floats64, scaling numerical features to standardize the data, and exporting categorical features in a vector format, similar to the OneHotEncoder, to transform categorical features into binary vectors. An example of vectorization of categorical data is for position. The position of the marker can be either alone, initial, medial, or final, respectively coded as 0, 1, 2 and 3. The vectorization transforms the value of 0-3 into a vector of length 4 with binary values. For instance, 0 is replaced by [1, 0, 0, 0] and 3 is replaced by [0, 0, 0, 1].

Different datasets are used in this chapter, explained in details in sections 10.3.1 and 10.4.1. All data for this chapter excludes spurts labeled with undetermined stance (x). In sum, the data for the first experiment mainly consists of lexical data from the spurts to test whether *um* and *uh* are relevant word unigrams for the classification of stance. The data for the second experiment consists of information on the position and the acoustic characteristics of the two markers, to test whether discourse and acoustic information improve accuracy over lexical information.

10.3 Experiment 1 : Classification of stance using lexical features

The goal of this section if to find out whether filtering *um* and *uh* from the spurts decreases the performance of the classification task, which is to predict stance labels using three stance distinctions, the presence of stance (binary stance strength), three-way stance strength, and stance polarity.

10.3.1 Data

Four datasets are used in this experiment : the first dataset contains lexical features with no *um* and *uh*, for a total of 8,984 spurts. This means that the two markers are filtered from the spurts and that any other word is left as is. The second dataset has no *um* and has 9,047 spurts, the third dataset has no *uh* and has 9,140 spurts, and finally, the fourth dataset contains both *um* and *uh* and has 9,203 spurts. The reason why the four datasets have different numbers of samples is because some spurts only consist of one *um* or one *uh*, which means that the spurt is entirely removed when the marker is filtered.

10.3.2 Methodology

Five features are exploited in this experiment, bag of words (*b.o.w.*), the top 200 most frequent words (*top 200*), the gender of the speaker (*gender*), the task (Budget Task vs. Inventory Task, see section 5.3.1 for more information) (*task*), and the duration of the spurt in seconds (*duration*). The bag of words and the top 200 features consist of word unigram features from the spurts (i.e., the utterance). I use the top 200 most common words to reduce the number of word unigrams, and to optimize runtime. The 200 cutoff was chosen to cover lexical words in addition to function words that have a high token frequency (see Appendix C). Furthermore, the feature *b.o.w.* does not improve the system's performance over the feature *top 200*, and it has a slower run time than *top 200*. Hence, *top 200* is used instead *b.o.w.* for the rest of the experiment.

Four feature combinations are used. The feature *no lex* means that no lexical features are taken into account (i.e., no *b.o.w.* or no *top 200*), only *gender*, *task*, and *duration* are used. The feature *all* stands for *top 200, gender, task,* and *duration.*

To deal with unbalanced classes in the four datasets mentioned in section 10.3.1, weights are added by using the function **class_weight** which takes either a dictionary of key classes and value weights, or the "balanced" value, which adjusts weights inversely proportional to the class frequency. The weights are chosen based on the initial distribution of the data and on the best precision, recall, F1-score, and accuracy scores (see section D.2 in Appendix D for the code).

Various measures are used to analyze the system's performance : accuracy scores, precision, recall, and F1-scores. The accuracy scores reported in the experiment are averaged over five folds, and the precision, recall, and F1-scores are taken from the fifth fold of each classification task for conciseness and consistency sake. Precision, recall, and F1-scores are used in addition to accuracy for several reasons. These performance measures are especially relevant when dealing with uneven classes. Furthermore, precision shows the positive predictive value (i.e., the classifier's exactness) and recall shows the true positive rate (i.e., the

classifier's completeness), while F1-scores inform us on the balance between precision and recall. These performance measures also inform us on how the algorithm performs on each class as opposed to its overall performance. For instance, the precision, recall, and F1-scores show that the feature combination *no lex* fails to assign one class in the prediction of binary stance and stance polarity, despite using class weights, and show that using lexical features (*all* or *top 200*) increases the performance of the classifier.

10.3.3 Results

Table 10.1 summarizes the baselines of the classifiers for the four datasets and for the three stance measures, based on the distribution of the data. These baselines are used for two reasons. First, to determine the weights assigned to the minority classes to avoid the algorithm to be biased towards the most common class. Second, to compare the accuracy of the classifiers in order to see which combination of features shows more improvement over the baseline levels.

TABLE 10.1: Baselines for the four datasets and the three stance measures, binary stance strength (0=no stance, 1=any stance), three-way stance strength (0=no stance, 1=weak stance, 2=moderate/strong stance), and stance polarity (-=negative, 0=neutral, +=positive)

Dataset	1 without *um* and *uh*		
Binary stance	0 : 22%	1 : 78%	
Three-way stance	0 : 22%	1 : 51%	2 : 27%
Stance polarity	- : 5%	0 : 65%	+ : 30%
Dataset	2 without *um*		
Binary stance	0 : 22%	1 : 78%	
Three-way stance	0 : 22%	1 : 51%	2 : 27%
Stance polarity	- : 5%	0 : 65%	+ : 30%
Dataset	3 without *uh*		
Binary stance	0 : 23%	1 : 77%	
Three-way stance	0 : 23%	1 : 51%	2 : 26%
Stance polarity	- : 5%	0 : 65%	+ : 30%
Dataset	4 with *um* and *uh*		
Binary stance	0 : 24%	1 : 76%	
Three-way stance	0 : 24%	1 : 50%	2 : 26%
Stance polarity	- : 5%	0 : 65%	+ : 30%

Table 10.2 summarizes the average accuracy scores over the five folds, the precision, recall, and F1-scores, in predicting binary stance, for the four datasets and for three feature combinations. Because the distribution of the four datasets is very similar in terms of binary stance, the same weights were used for this classification task : .6 for class 0 (no stance) and .4 for class 1 (the presence of stance). The weights were assigned using the following func-

tion : **wclf_2 = SVC(kernel='linear', class_weight=0 :.60, 1 :.40, gamma=0.001, C=1.0)**, to balance the uneven classes listed in Table 10.1.

All performance measures, especially precision, recall, and F1 scores show that the feature combination *no lex* (*gender, duration,* and *task*) does not add information to the model, and that it leads to accuracy scores below baseline for datasets 1 and 2, and barely above baseline for datasets 3 and 4. Furthermore, the results for *top 200* and *all* show similar performances, which means that the lexical features account for most of the information in predicting binary stance in this corpus.

Based on the baselines listed in Table 10.1, the biggest improvement over baseline is for dataset 4 (i.e., including *um* and *uh*) with an accuracy increase of 6.27% from 76% to 80.77% by using the lexical feature *top 200*. The least improvement in accuracy is for dataset 1 (i.e., *um* and *uh* filtered) with an increase of 3.12% from 78% to 80.44%. These results suggest that the best predictions are achieved by not filtering *um* and *uh*, indicating that the two markers are valuable word unigrams in the prediction of binary stance in this corpus. It is important to keep in mind however that the difference in performance between the two datasets is rather small, and that further investigation should shed light on the significance of the difference between using and filtering the features *um* and *uh* when predicting stance.

The comparison of the system's performance (accuracy, precision, recall, and F1-scores) between datasets 2 and 3 informs us on whether filtering *um* vs. *uh* decreases the performance of the system, and allows comparing the importance of each marker as a word unigram feature. In dataset 3 where *uh* is filtered, the accuracy scores show an increase of 4.77% over baseline, from 77% to 80.68% and 80.67% respectively for the features *top 200* and *all*. In dataset 2 where *um* is filtered, the performance increases by 3.15% over baseline, from 78% to 80.46%, using the *top 200* feature. These results indicate that the system performs worse when *um* is filtered than when *uh* is filtered, which means that *um* is likely to play a slightly more important role than *uh* in predicting binary stance in this corpus.

TABLE 10.2: Summary of mean accuracy sores for the five folds (M. acc.), Precision (P.), Recall (R.), and F1-scores (F1) for classes 0 and 1, for the four datasets, for three feature combinations (*no lex*, *top 200*, and *all*), in predicting binary stance (i.e., no stance : 0 or the presence of stance : 1)

Features	M. acc.	P. 0	P. 1	R. 0	R. 1	F1 0	F1 1
Dataset	1 without *um* and *uh*						
no lex	77.99	0	0.771	0	1	0	0.871
top 200	80.43	0.540	0.875	0.589	0.851	0.563	0.863
all	80.44	0.541	0.875	0.591	0.851	0.565	0.863
Dataset	2 without *um*						
no lex	77.56	0	0.763	0	1	0	0.865
top 200	80.46	0.558	0.878	0.620	0.847	0.587	0.862
all	80.39	0.553	0.875	0.613	0.846	0.581	0.860
Dataset	3 without *uh*						
no lex	77.73	0	0.760	0	1	0	0.864
top 200	80.68	0.568	0.875	0.615	0.852	0.591	0.864
all	80.67	0.568	0.875	0.615	0.853	0.591	0.864
Dataset	4 with *um* and *uh*						
no lex	76.32	0	0.752	0	1	0	0.858
top 200	80.77	0.582	0.880	0.650	0.846	0.614	0.863
all	80.74	0.580	0.879	0.648	0.845	0.612	0.862

Table 10.3 summarizes the mean accuracy scores over five folds, the precision, recall, and F1-scores, for classes 0 (no stance), 1 (weak stance), and 2 (moderate/strong stance), for the automatic prediction of three-way stance strength. The weights used in this task are based on the baselines listed in Table 10.1 : .4 for class 0, .2 for class 1, and .4 for class 2 ; using

the following function : **wclf_3 = SVC(kernel='linear', class_weight=0 :.4, 1 :.2, 2 :.4, gamma=0.001, C=1.0)**). Since the four datasets have a similar distribution across the three classes the same weights are used across all datasets.

Similarly to binary stance, the accuracy scores indicate that the best predictions are obtained by using dataset 4 where *um* and *uh* are not filtered. The best results show an improvement of 19.92% over baseline, from 50% to 59.96%, by using the feature *top 200*. Dataset 1 shows the least improvement with an increase of 15.92%, from 51% to 59.12%, by using the feature *all*. However, it is interesting to note that the F1-scores indicate that the system performs better for classes 1 and 2 in dataset 1 than in dataset 4, but not for class 0. These performance metrics further indicate that the difference in performance between the two datasets is very small, and that it is important to consider different metrics when looking at a system's performance, especially when dealing with unbalanced classes.

The results from datasets 2 and 3 show that filtering *um* leads to slightly lower improvement (15.92% increase over baseline, from 51% to 59.12%) than filtering *uh* (17.08% increase over baseline, from 51% to 59.71%). These results indicate that *um* is likely to play a slightly more important role in predicting three-way stance than *uh*, although the differences are small.

Unlike for binary stance, these results also indicate that the feature combination *all* leads to slightly better results than the feature *top 200* in the first three datasets. However, in dataset 4 both *all* and *top 200* show improvement depending on the performance metric. The feature *top 200* leads to a slightly higher accuracy score while the feature combination *all* leads to slightly higher F1-scores for 2 out of 3 classes. Finally, the performance scores, especially accuracy and precision, show that the absence of lexical features (*no lex*) fails to predict three-way stance strength, showing accuracy scores below baseline.

TABLE 10.3: Summary of mean accuracy sores for the five folds (M. acc.), Precision (P.), Recall (R.), and F1-scores (F1) for classes 0, 1, and 2, for the four datasets, for three feature combinations (*no lex*, *top 200*, and *all*), in predicting three-way stance strength (i.e., no stance : 0, weak stance : 1, and moderate/strong stance : 2)

Features	M. acc.	P. 0	P. 1	P. 2	R. 0	R. 1	R. 2	F1 0	F1 1	F1 2
Dataset	1 without *um* and *uh*									
no lex	47.46	0.275	0.599	0.502	0.421	0.473	0.484	0.333	0.529	0.493
top 200	58.79	0.468	0.757	0.515	0.742	0.554	0.535	0.574	0.640	0.525
all	59.12	0.466	0.751	0.548	0.752	0.576	0.512	0.575	0.652	0.529
Dataset	2 without *um*									
no lex	47.55	0.288	0.594	0.502	0.427	0.476	0.482	0.344	0.529	0.492
top 200	58.87	0.485	0.751	0.514	0.751	0.558	0.527	0.589	0.640	0.520
all	59.12	0.480	0.749	0.540	0.762	0.573	0.506	0.589	0.649	0.523
Dataset	3 without *uh*									
no lex	47.60	0.297	0.594	0.504	0.442	0.473	0.482	0.355	0.527	0.493
top 200	59.42	0.483	0.753	0.510	0.759	0.548	0.525	0.590	0.634	0.517
all	59.71	0.481	0.751	0.549	0.778	0.571	0.511	0.592	0.649	0.529
Dataset	4 with *um* and *uh*									
no lex	47.63	0.310	0.588	0.505	0.446	0.475	0.483	0.366	0.526	0.494
top 200	59.96	0.501	0.756	0.513	0.770	0.553	0.528	0.607	0.639	0.520
all	59.85	0.495	0.749	0.544	0.775	0.568	0.509	0.604	0.646	0.526

Table 10.4 summarizes the performance of the classifier in predicting stance polarity. Since the four datasets have the same distribution, the same weights are used : .55 for negative stance, .15 for neutral stance, and .3 for positive stance ; using the following function : (**wclf_4 = SVC(kernel='linear', class_weight=-1 :.55, 0 :.15, 1 :.3, gamma=0.001,**

C=1.0)).

Similarly to the prediction of binary stance, the precision, recall, and F1-scores indicate that the feature combination *no lex* fails to assign the minority class (class -1). The lexical feature (*top 200*) or the combination of all features (*all*) show improvement in terms of performance over baseline. The accuracy scores indicate that the combination *all* leads to the highest performance in datasets 1, 2, and 4, but the precision, recall and F1-scores show that the feature *top 200* leads to overall better results, especially for the minority class, therefore indicating that *top 200* leads to better overall performance of the system in predicting stance polarity in this corpus.

The accuracy scores show that the biggest improvement is obtained by using dataset 4 with the feature combination *all*, showing an increase of 16.87% over baseline, from 65% to 75.97%. The lowest performance increase is obtained in dataset 1, with an increase of 16.07% over baseline, from 65% to 75.45%. Similarly to binary and three-way stance strength, these results indicate that using the features *um* and *uh* leads to a slight increase in performance.

The comparison of the system's performance in datasets 2 and 3 also indicates that filtering *um* (16.26% increase over baseline, from 65% to 75.57% with the feature *top 200*) leads to slightly less improvement than filtering *uh* (16.69% increase, from 65% to 75.85% with the same feature), although the difference is very slight.

TABLE 10.4: Summary of mean accuracy sores for the five folds (M. acc.), Precision (P.), Recall (R.), and F1-scores (F1) for classes -1, 0, and +1, for the four datasets, for three feature combinations (*no lex*, *top 200*, and *all*), in predicting stance polarity (i.e., negative : -1, neutral : 0, and positive : +1)

Features	M. acc.	P. -1	P. 0	P. +1	R. -1	R. 0	R. +1	F1 -1	F1 0	F1 +1
Dataset	1 without *um* and *uh*									
no lex	47.48	0	0.736	0.335	0	0.275	0.865	0	0.400	0.483
top 200	75.42	0.325	0.877	0.624	0.283	0.740	0.857	0.302	0.803	0.722
all	75.45	0.314	0.875	0.619	0.239	0.741	0.857	0.272	0.803	0.719
Dataset	2 without *um*									
no lex	47.98	0	0.672	0.401	0	0.490	0.661	0	0.567	0.499
top 200	75.57	0.325	0.881	0.625	0.283	0.746	0.858	0.302	0.808	0.723
all	75.58	0.314	0.879	0.620	0.239	0.748	0.858	0.272	0.808	0.720
Dataset	3 without *uh*									
no lex	49.49	0	0.666	0.306	0	0.321	0.724	0	0.434	0.431
top 200	75.85	0.329	0.881	0.625	0.283	0.748	0.857	0.304	0.809	0.723
all	75.83	0.314	0.879	0.620	0.239	0.749	0.857	0.272	0.809	0.720
Dataset	4 with *um* and *uh*									
no lex	48.76	0	0.656	0.293	0	0.288	0.730	0	0.400	0.419
top 200	75.94	0.321	0.885	0.627	0.283	0.753	0.859	0.301	0.814	0.725
all	75.97	0.310	0.883	0.623	0.239	0.755	0.859	0.270	0.814	0.722

10.3.4 Summary

The performance metrics presented in this experiment, whether looking at accuracy, precision, recall, or F1-scores, show that the combination of all features (*all*) and the top

200 most common word unigram features (*top 200*) lead to the highest performance for each of the three stance predictions. These results indicate that the lexical features therefore account for all of the information in the model, and that *task, gender*, and *duration* do not add information to the model. This is further corroborated by the fact that the feature combination *no lex* fails to predict the three stance dimensions, indicated by accuracy scores below or barely above baseline, and by the precision, recall, and F1-scores revealing low performance on the minority class despite the use of weights.

The performance metrics also show that the lexical features used in this experiment show the most improvement over baseline when predicting three-way stance strength (19.92% improvement over baseline), followed by stance polarity (16.87%), and finally show the least improvement when predicting binary stance strength (4.77%). These results indicate that lexical features are most informative when predicting three-way stance strength and stance polarity.

For each of the stance predictions, the best results are obtained by using dataset 4 which contains both *um* and *uh*, indicating that *um* and *uh* are relevant word unigram features. However, it is important to keep in mind that those differences are rather small, especially when predicting stance polarity. These results suggest that *um* and *uh* might play a less important role as word unigram features compared to other words in predicting stance polarity than in predicting binary and three-way stance strength in this corpus. Further research on the topic will investigate the significance of the difference between using and filtering the two markers.

Finally, the comparison of the performance of the algorithms in datasets 2 and 3 shows that filtering *um* leads to slightly lower performances than filtering *uh* for each of the three stance predictions, therefore suggesting that *um* plays a slightly more important role than *uh* in predicting stance in this corpus. However, given the small differences in performance between the two markers, this finding requires further investigation to confirm the relative role of *um* and *uh* in stance prediction.

10.4 Experiment 2 : Classification of stance using acoustic and discourse features

The goal of this experiment is to test whether acoustic cues and the position of the two markers can predict stance labels, and whether these cues improve accuracy over lexical cues tested in the previous experiment.

10.4.1 Data

Three datasets are used in this experiment : dataset 1 contains information on *um* and *uh*, dataset 2 contains information on *um* only, and dataset 3 contains information on *uh* only. Similarly to the previous experiment, datasets 2 and 3 are used to compare the relative contribution of *um* and *uh* in predicting stance. Dataset 1 has 984 samples, dataset 2 has 568 samples, and dataset 3 contains 416 samples. Each sample of the data consists of a list of features providing information on the markers such as their position and their acoustic properties.

10.4.2 Methodology

A total of 11 features is exploited in this analysis, including acoustic characteristics (duration, f1, f2, and intensity), position (alone, initial, medial, or final), as well as word (*um* vs. *uh*), gender (man vs. woman), and task (Inventory vs. Budget). The word feature is only used in dataset 1 containing *um* and *uh*. Pitch is excluded from this analysis because there are 453 missing values out of 984 samples (46%) due pitch extraction errors. Solutions to deal with missing values include removing missing values, or replacing them with the median. I did not remove the samples with the missing values because they represent nearly half of the data, and I did not want to replace the values with the median because the missing values are likely on either end of the pitch range. I therefore excluded the entire pitch feature in order to keep the 984 samples.

Several feature combinations are used in this experiment, ranging from excluding one

feature to a combination of features, and only the most relevant ones are discussed in the results section. The feature *all* stands for all 11 features. *No position* stands for all features but the position features (i.e., alone, initial, medial, and final are excluded), *no alone/init* means that the position features alone and initial are excluded, *no med/final* means that the medial and final position features are excluded, and *no alone/init/med* means that only the final position feature is included. *No acoustics* stands for all features but the acoustic features (i.e., duration, f1, f2, and intensity are excluded). Finally, the feature *no word* means that all features are taken into account, except for the marker (i.e., *um* vs. *uh*).

10.4.3 Results

Table 10.5 summarizes the baselines for the three datasets used in this experiment and for the three stance measures, based on the distribution of the classes. The baselines listed here are used to compare the performance of the classifiers in this experiment to show improvement over baseline as opposed to raw accuracy scores. The baselines indicate that each dataset has its own distribution and minority classes, and that the weights need to be adjusted for each dataset and each stance dimension.

TABLE 10.5: Baselines for the three datasets and the three stance measures, binary stance strength (0=no stance, 1=any stance), three-way stance strength (0=no stance, 1=weak stance, 2=moderate/strong stance), and stance polarity (-=negative, 0=neutral, +=positive)

Dataset	1 *um* and *uh*		
Binary stance	0 : 40%	1 : 60%	
Three-way stance	0 : 40%	1 : 30%	2 : 30%
Stance polarity	- : 6%	0 : 80%	+ : 14%
Dataset	2 *um*		
Binary stance	0 : 46%	1 : 54%	
Three-way stance	0 : 46%	1 : 29%	2 : 25%
Stance polarity	- : 3%	0 : 82%	+ : 15%
Dataset	3 *uh*		
Binary stance	0 : 31%	1 : 69%	
Three-way stance	0 : 31%	1 : 33%	2 : 37%
Stance polarity	- : 12%	0 : 78%	+ : 10%

The main results of the performance of the classifiers for the three datasets are summarized in Table 10.6. Because all datasets have similar trends, various feature combinations are illustrated in different datasets for clarity purposes. The weights used in this classification task for classes 0 and 1 respectively are : .6 and .4 for dataset 1, .55 and .45 for dataset 2, and .7 and .3 for dataset 3.

The results show that across the three datasets, the position features account for all predictions in the model, as illustrated by the results for the combination of all features (*all*) and the combination of position features alone (*position*). Results not listed in this table show that removing the word feature (*um* vs. *uh*) and removing one position feature at a time does not affect the performance of the system for any of the datasets. However, results illustrated

in dataset 1 show that removing all position features decreases the accuracy of the algorithm below baseline, and that removing the combination of all position features but the final feature (*no alone/init/med*) or removing the features *alone* and *initial* (*no alone/init*) also decreases the system's performance compared to including all position features. However, removing the feature combination *medial* and *final* (*no med/final*) does not decrease the system's performance. These findings indicate that the position features are not informative by themselves, and that the more position features are included, the more the performance increases.

Removing all acoustic features (*no acoustics*) or one acoustic feature at a time does not show any decrease in the system's performance. The relative contribution of each acoustic feature is analyzed by including one feature at a time. The performance metrics show performance below baseline and precision and recall scores equal to 0 for most models. The performance metrics show slight improvement when using f2 for *um*, and a slight bump in accuracy score when using duration for *uh*. These results indicate that the acoustic features do not bring any information over the position features, and that they account for close to no information.

The precision and recall scores also show that in datasets 1 and 3 the recall of the minority class (0 : no stance) is very low, especially in dataset 3, which means that the algorithm has a low positive rate for this class. These results also show that the position features show an improvement in accuracy over baseline of 31.76% in dataset 1, 45.74% in dataset 2, and 15.31% in dataset 3. These results indicate that the position features are relevant features when looking at *um* and *uh* to predict binary stance, and that the position of *um* is more informative than the position of *uh* in this corpus.

TABLE 10.6: Summary of mean accuracy sores for the five folds (M. acc.), Precision (P.), Recall (R.), and F1-scores (F1) for classes 0 and 1, for the three datasets, for various feature combinations, in predicting binary stance (i.e., no stance : 0 or the presence of stance : 1)

Features	M. acc.	P. 0	P. 1	R. 0	R. 1	F1 0	F1 1
Dataset	1 *um* and *uh*						
no position	56	0.493	0.705	0.465	0.728	0.478	0.717
no alone/init/med	63.11	0.534	0.710	0.437	0.784	0.481	0.745
no alone/init	72.36	0.626	0.867	0.803	0.728	0.704	0.791
no med/final	79.06	1	0.776	0.493	1	0.660	0.874
no acoustics	79.06	1	0.776	0.493	1	0.660	0.874
position	79.06	1	0.776	0.493	1	0.660	0.874
all	79.06	1	0.776	0.493	1	0.660	0.874
Dataset	2 *um*						
f1	44.9	0.469	0	1	0	0.639	0
intensity	45.07	0.469	0	1	0	0.639	0
duration	45.78	0.469	0	1	0	0.639	0
f2	46.65	0.4694	0.333	0.962	0.017	0.626	0.032
position	78.70	0.947	0.773	0.679	0.967	0.791	0.859
all	78.70	0.947	0.773	0.679	0.967	0.791	0.859
Dataset	3 *uh*						
f2	40.33	0.265	0	1	0	0.419	0
f1	40.33	0.278	1	1	0.066	0.436	0.123
intensity	42.02	0.265	0	1	0	0.419	0
duration	45.87	0.265	0	1	0	0.419	0
position	79.06	1	0.776	0.493	1	0.660	0.874
all	79.06	1	0.776	0.493	1	0.660	0.874

Table 10.7 summarizes the most relevant results of the classifiers in predicting three-way stance strength for the three datasets with various features combinations : all features but the position features (*no position*), the position features only (*postion*), all 11 features (*all*), removing the word feature only (*no word*), all features except the acoustic ones (*no acous.*), the acoustic features only (*acoustics*), duration only (*duration*), intensity only (*intensity*), and duration and intensity only (*dur + intensity*). The weights used for each dataset for classes 0 (no stance), 1 (weak stance), and 2 (moderate/strong stance) are : 0 :.26, 1 :.37, 2 :.37 for dataset 1, 0 :.23, 1 :.36, 2 :.41 for dataset 2, and 0 :.34, 1 :.33, 2 :.33 for dataset 3.

The performance measures (accuracy, precision, recall, and F1-scores) show that similarly to binary stance, the position features account for most of the predictions. In addition, the performance measures also show that removing the word feature (*no word*) and the acoustic features (*no acous.*) increases the performance of the system, which means that these features add noise to the model when using dataset 1. Further results not illustrated in Table 10.7 show that the accuracy scores drop when removing the task feature for three-way stance strength (e.g., from 56.4 to 49.88 for dataset 1), which is not surprising since the Budget Task elicits more speaker involvement (see section 5.3.1 for more information) than the Inventory task.

For datasets 2 and 3 the combination of all features leads to the highest performance with an improvement over baseline of 24.02%, from 46% to 57.05% for *um*, and of 51.32%, from 37% to 55.99% for *uh* whereas removing acoustic features does not affect the performance for dataset 2 and decreases the performance for dataset 3, especially for classes 1 and 2. It is also interesting to note that using acoustic features only leads to performance below baseline when using *um*, but not for *uh*. In fact, using the intensity or duration features only, or as a combination increases the accuracy of the system above baseline (37%), by 11.05% for intensity, 11.18% for duration, and 24.08% for duration and intensity. These results indicate that the acoustic information of the marker *uh* (especially duration and intensity) is more valuable than the acoustic information of the marker *um* when predicting three-way stance strength in this corpus. In sum, the performance of the different models summarized in this

section suggest that that features related to *uh* are more informative than those pertaining to *um* in the prediction of three-way stance strength in this corpus.

TABLE 10.7: Summary of mean accuracy sores for the five folds (M. acc.), Precision (P.), Recall (R.), and F1-scores (F1) for classes 0, 1, and 2, for the three datasets, for various feature combinations, to predict three-way stance strength (i.e., no stance : 0, weak stance : 1, and moderate/strong stance : 2)

Features	M. acc.	P. 0	P. 1	P. 2	R. 0	R. 1	R. 2	F1 0	F1 1	F1 2
Dataset	1 *um* and *uh*									
no position	43.5	0.481	0.350	0.529	0.352	0.292	0.714	0.407	0.318	0.308
position	54.57	1	0.384	0.648	0.493	0.583	0.740	0.660	0.463	0.691
all	56.4	0.9	0.4	0.721	0.507	0.792	0.571	0.649	0.531	0.638
no word	56.91	0.878	0.379	0.652	0.507	0.521	0.753	0.643	0.439	0.699
no acous.	57.82	1	0.391	0.641	0.493	0.562	0.766	0.620	0.462	0.698
Dataset	2 *um*									
acoustics	37.5	0.495	0.444	0	0.868	0.108	0	0.630	0.174	0
position	53.88	0.947	0.394	0.381	0.679	0.351	0.696	0.791	0.371	0.492
no acous.	57.05	0.818	0.533	0.393	0.849	0.432	0.478	0.833	0.478	0.431
all	57.05	0.818	0.533	0.393	0.849	0.432	0.478	0.833	0.478	0.431
Dataset	3 *uh*									
acoustics	40.62	0.286	0.296	0.592	0.091	0.400	0.707	0.138	0.340	644
intensity	41.09	0.250	0.325	0.355	0.103	0.464	0.423	0.146	0.382	0.386
duration	41.14	0.500	0.429	0.547	0.273	0.150	0.854	0.353	0.222	0.667
dur + intensity	45.91	0.625	0.368	0.589	0.227	0.350	0.805	0.333	0.359	0.680
no acous.	49.73	0.5	0.200	0.750	0.773	0.450	0.073	0.607	0.277	0.133
position	51.91	1	0.440	0.714	0.409	0.550	0.854	0.581	0.489	0.778
all	55.99	0.533	0.545	0.786	0.727	0.300	0.805	0.615	0.387	0.795

Spurts that contain *um* and *uh* are especially unevenly distributed across the three po-

larity categories. Figure 8.2 from Chapter 8 is repeated here for clarity purposes (Figure 10.1) [1]. Table 10.5 and the right-side plot of Figure 10.1 illustrate the distribution of the spurts that contain *um* and *uh* across polarity values, and shows that most markers are used in neutral spurts, and that very few *ums* are used in negative spurts. Weights are used in order to counterbalance predictions for the majority class (i.e., neutral spurts) and lead to performance scores well below baseline. Further research will aim at investigating the role of features pertaining to *um* and *uh* to predict stance polarity.

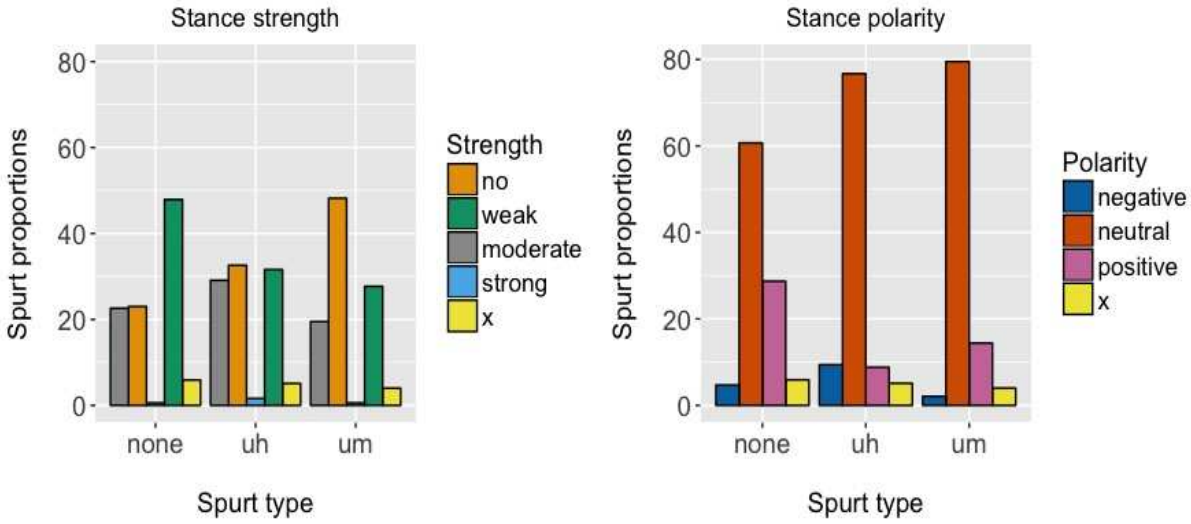

FIGURE 10.1: Proportions of spurts for stance strength (left) and stance polarity (right) for each spurt type

10.5 Summary

The findings from this experiment show that acoustic features do not increase accuracy when predicting binary stance using *um* and/or *uh* and when predicting three-way stance strength using *um*. However, when looking at *uh* to predict three-way stance strength, the model with just acoustic features shows an improvement of 9.78% over baseline, and a 24.08%

1. Note that all spurts marked with undetermined stance (x) are excluded in this experiment.

improvement when using duration and intensity only. These results indicate that the acoustic characteristics of *uh*, especially duration and intensity, help predicting three-way stance strength in this corpus. Further more, these results are consistent with results from Chapter 9 which suggest that three-way stance is a better strength categorization because it captures the differences between weak and moderate-to-strong stance.

Omitting position features for both binary and three-way stance strength leads to the biggest drop in accuracy in all datasets, and results show that the marker's position is the most informative feature in the prediction of binary and three-way stance strength. It is also interesting to note that the word feature (*um* vs. *uh*) does not seem to impact the system's performance, which means that whether the marker is *um* or *uh* does not add information to the model. Finally, the comparison of the performances for datasets 2 and 3 show that features pertaining to *um* are more informative than those pertaining to *uh* to predict binary stance, while features pertaining to *uh* are more informative to predict three-way stance strength.

10.6 Chapter summary and discussion

The findings from section 10.3 indicate that incorporating *um* and *uh* as word unigram features into a stance classification model increases the performance of the algorithm, especially when predicting three-way stance strength and stance polarity. The findings also suggest that *um* plays a slightly more important role than *uh* for each of the three stance predictions. These results are corroborated by findings from Levow et al. (2014) that report *um* is a key lexical feature in automatic stance strength recognition. However, it is important to keep in mind that the differences reported in this experiment in performance between using and filtering *um* and *uh* are small, and that the relative contribution of *um* and *uh* as word unigram features needs to be further investigated to find out whether the difference is significant.

The findings from section 10.4.1 show that the position features of *um* and *uh* are the most important features in predicting binary stance strength, especially for *um*, and in predicting three-way stance strength. The results from this experiment also show that the acoustic

features of *um* do not add information to the model, while the acoustic features of *uh* do, especially duration and intensity.

The findings from the two experiments conducted in this chapter show that features pertaining to the two markers play an role in predicting stance in this corpus : *um* and *uh* are relevant word unigram features and their position and the acoustic properties of *uh* lead to increases in the system's performance. To summarize, *um* seems to play a slightly more important role than *uh* as a word unigram feature when predicting all three stance dimensions. Features pertaining to *um* play a more important role than those pertaining to *uh* to predict binary stance, and finally, features pertaining to *uh* play a more important role than those pertaining to *um* when predicting three-way stance strength.

Furthermore, the fact that the models predicting three-way stance strength lead to greater improvements over baseline in each experiment is corroborated by results from Chapters 8 and 9 which point out that three-way stance is a more accurate distinction because there is a lot of variation between weak and moderate-to-strong stance in terms of discourse and acoustic properties of the markers.

Chapter 11
CONCLUSION

This chapter summarizes the findings of each chapter, discusses the contributions and the impact of this study, and examines future directions for this research on the two markers *um* and *uh*.

11.1 Result summary

The sum of the results confirms that *um* and *uh* are distinct entities and that they are not random (i.e., they play an informational role). These findings corroborate findings from previous studies and show that they extend to the ATAROS corpus. One of the main contributions on this topic is that *um* and *uh* are subject to more misperceptions than function words and other frequent words, and that *uh* undergoes more transcription errors than *um*, suggesting that *um* is more salient and that it might play a more important role in discourse than *uh*. Anther important contribution of this study is that it provides evidence in favor of the association between the two markers and stance marking in ATAROS, and it sheds light on the relative contribution of the two markers as discourse and stance markers.

The results from Chapter 6 on *um* and *uh* serve as a baseline for further analyses on the two corpora and on the two markers. The analysis shows that *um* and *uh* have different distributions and different duration cues, which separates them. The two markers should therefore be treated as separate entities, and further research should consider the effect of gender, naturalness, and speaker involvement when looking at *um* and *uh*. The results also show that factors such as gender, speaker involvement, and naturalness affect *um* more than *uh*, and show that the two markers behave differently in the two speech activities investigated in this chapter. Future research will aim at further investigating the relative role of *um* and

uh as discourse markers, and their role across various speech activities.

In Chapter 7, the analyses of the production of *um* and *uh* and of the transcription errors show again that *um* and *uh* are affected by most discourse and metalinguistic variables, and that they generally behave differently from other words and from each other. Transcribers miss and hallucinate *uh* more than *um*, and variables such as naturalness of the conversation or difficulty to transcribe reliably affect the transcription errors of *um* and *uh*, and in different trends. Variables related to the speaker such as gender, age, dialect, or education, have a different effect on the production of the two markers. Additional results on word type show that *um* and *uh*, as well as words from the same category (i.e., *other*), are proportionally more missed and hallucinated than other frequent words such as function words. These findings corroborate the fact that *um* and *uh* are different since they have different distributions and different transcription error rates. Finally, the fact that *um* has less transcription errors than *uh* indicates that *um* is more likely to carry a higher informational load in discourse than *uh*, providing evidence that *um* might play a more important role in discourse than *uh*.

The findings in Chapter 8 show there is a relationship between *um* and *uh* and the stance of the spurt in which they are found. The results show that the presence and position of the markers, as well as the stance labels, are dependent variables. Spurts that contain one *um*, one *uh*, or no markers behave differently from each other, and spurts are more likely to be marked with a certain degree or polarity of stance depending on the position of the marker. For instance, isolated *ums* and *uhs* are more likely to be marked with no stance and neutral polarity, while spurts that contain markers in other positions are more likely to be marked with some degree of stance, as well as negative or positive polarity. These findings lend support for the fact that *um* and *uh* are associated with stance marking, and that their presence and position can be used as a predictor for the stance label of the spurt.

The statistical analyses from Chapter 9 do not find reliable effects for stance on the acoustic realization of the vowel in *um* and *uh*. However, the differences in the medians indicate that he presence of stance (i.e., for binary stance) correlates with shorter vowels, and it affects vowel pitch, intensity, and quality in different ways for *um* and *uh*. Three-way

stance affects *uh* and *other* in similar ways, and differently from *um*, for pitch, intensity, and F2. Three-way stance affects F1 differently in all word groups. Similarly, polarity affects the pitch and the intensity of the vowels in *uh* and *other* in similar ways, while *um* behaves differently. These results show that the acoustic realization of *um* and *uh* varies depending on stance, and that pitch and intensity follow similar patterns.

In Chapter 10, I incorporate findings from the previous experiments in a concrete application, which consists in using information pertaining to *um* and *uh* to train classifiers to automatically label stance. The results from the first classification experiment reveal that *um* and *uh* are important word unigram features in the prediction of stance, that *um* seems to be a more important feature than *uh*, and that lexical features best predict three-way stance strength, followed by stance polarity and binary stance. In the second experiment, the performances in classifying stance using discourse and acoustic features pertaining to *um* and *uh* show that different acoustic features carry different levels of information depending on the marker and on the stance classification. Position is the most important feature for all models and acoustic features (duration and intensity) only increase the performance when using *uh* to predict three-way stance strength. The results also show that using *um* leads to better results when predicting binary stance, while using *uh* leads to better results when predicting three-way stance strength. These results are consistent with results from other chapters as well as other studies.

In sum, the findings of this dissertation show that *um* and *uh* are different entities and indicate that *um* and *uh* play different roles, especially with regards to stance predictions. The discourse and acoustic features of *um* and *uh* are different. The marker *um* varies to a greater extent than the marker *uh*. Transcribers perceive *um* better than *uh*. The acoustic properties of *um* and *uh* vary depending on stance strength and polarity. The word unigram feature *um* seems to play a more important role than *uh* to predict stance. Features associated to *um* increase accuracy of automatic stance classification over features associated to *uh* to predict binary stress, and features associated to *uh* increase accuracy over those associated to *um* to predict three-way stance strength. The work presented in this dissertation provides

support to show *um* and *uh* are not just fillers or disfluencies, but rather that they have a wide range of uses, from fillers to pragmatic and stance markers.

11.2 Contributions and impact

The experiments from this dissertation provide a global analysis of *um* and *uh* in spontaneous speech using two corpora, consisting of different speech activities. One of the main contributions of this study is that it looks at how *um* and *uh* pattern depending on several variables, encompassing metalinguistic, semantico-discursive, and acoustic approaches, to provide a holistic understanding of how the two markers are used in spontaneous speech.

This study also raises issues about cross-corpora studies, such as how to compare results across different speech activities and across different annotation systems. Some of the main issues include segmenting speech into meaningful units for discourse and syntactic analysis, disfluency annotations, and fine grained stance annotations. A more universal annotation framework of speech corpora would therefore provide more access to cross-corpora studies, to increase the robustness of results, and to increase our understanding of the language mechanisms of different speech activities.

Another important contribution to the understanding of markers such as *um* and *uh* is the analysis of the production and the misperception of *um* and *uh* in the Switchboard corpus, compared to any other word, and compared to similar words with comparable frequency. Among other things, this analysis shows that the two markers behave differently from each other in terms of production, and that their rate of transcription error is not random.

Furthermore, as mentioned in Chapters 2 and 3, several state of the art language models still consider *um* and *uh* as mere filled pauses or speech disfluencies, and filter them from the signal. Possible applications of the findings from this study therefore include implementations of the roles of *um* and *uh* in language models, to improve general spontaneous speech processing and understanding, as well as automatic stance recognition. Furthermore, as we learn more about how discourse markers are used in different settings, and as we get a deeper understanding of how *um* and *uh* are used in various spontaneous speech activities (e.g.,

negotiation, collaboration, conversation, story telling, etc.), we could use that knowledge to build speech task or speech activity detectors, depending on how speakers use markers such as *um* and *uh*.

11.3 Future directions

This dissertation adopts a systematic and multidimensional analysis of *um* and *uh* in spontaneous speech. However, given the scope of the possible analyses on the two markers, there are still several dimensions to explore in order to gain a better understanding of how speakers use *um* and *uh*. Future work will expend on the prosodic environments in which the markers are used, in order to better understand their properties relative to prosodic units. Questions of interest include the association of the markers with pitch fluctuations compared to neighboring environments and to the speaker's pitch range. Other acoustic features of the markers should also be explored, such as the third formant, phonation types, and better pitch measures. Finally, in addition to vowel duration, future work will also investigate marker duration.

Further analyses of *um* and *uh* will also take into account how their production relates to syntactic structures and complexity, to find out whether they correlate with certain constituent types, structures, and properties. These studies will include variables such as constituent length, type, place, the constituents above and under, and the number of constituents in the sentence.

Future work will also explore the perception and the misperception of the two markers by investigating whether listeners always perceive the markers in the same way depending on their position in the prosodic unit, the syntactic constitutent in which they are, neighboring words, the acoustic realization of the marker, and depending on the stance value and the speech task.

Future work will also include more in depth understanding of the sublexical properties of the markers in order to find out the different meanings of *um* and *uh* from a semantic and pragmatic point of you. Ultimately, the goal is to better understand how the markers are

used, and whether we can find out automatic ways to identify the functions of the marker in spontaneous speech. In other words, can we automatically identify which markers are filled pauses, disfluencies, discourse markers, interjections, stance markers, and other possible uses of *um* and *uh* in various speech activities.

Bibliographie

Adda-Decker, M., Habert, B., Barras, C., Adda, G., Boula de Mareüil, P., and Paroublek, P. (2004). Une étude des disfluences pour la transcription automatique de la parole spontanée et l'amélioration des modèles de langage. In *Actes des 25emes Journées d'Etudes sur la Parole (JEP)*.

American Speech-Language-Hearing Association Ad Hoc Committee on Service Delivery in the Schools (1993). Definitions of communication disorders and variations. *ASHA*, 35(Suppl. 10) :40–41.

Anderson, N. B. and Shames, G. H. (2011). *Human Communication Disorders : An Introduction*. Pearson/Allyn and Bacon, 8th edition.

Arnold, J. E., Kam, C. L. H., and Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard : the on-line attribution of disfluency during reference comprehension. *Journal of experimental psychology. Learning, memory, and cognition*, 33(c) :914–930.

Arnold, J. E. and Tanenhaus, M. K. (2011). Disfluency Effects in Comprehension : How New Information Can Become Accessible. In *The Processing and Acquisition of Reference*, pages 197–217. MIT Press.

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., and Fagnano, M. (2004). The Old and Thee, uh, New. *Psychological Science*, 15(9) :578–582.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1) :1–48.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman, Harlow, England ; [New York].

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Blomgren, M., Nagarajan, S. S., Lee, J. N., Li, T., and Alvord, L. (2003). Preliminary results of a functional MRI study of brain activation patterns in stuttering and nonstuttering speakers during a lexical access task. *Journal of Fluency Disorders*, 28(4) :337–356.

Boersma, P. and Weenink, D. (2015). Praat : doing phonetics by computer.

Brown, S. F. (1945). The Loci of Stutterings In The Speech Sequence. *The Journal of Speech Disorders*, 10(3) :181–192.

Calhoun, S., Carletta, J., Brenier, J. M., Mayo, N., Jurafsky, D., Steedman, M., and Beaver, D. (2010). The NXT-format Switchboard Corpus : a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44(4) :387–419.

Chomsky, N. (1965). *Aspects of the theory of syntax*. M.I.T. Press, Cambridge.

Clark, H. H. and Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84(1) :73–111.

Corley, M. and Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech : The meaning of um. *Linguistics and Language Compass*, 2 :589–602.

Daly, D. A. and Burnett, M. L. (1996). Cluttering : Assessment, treatment planning, and case study illustration. *Journal of Fluency Disorders*, 21(3-4) :239–248.

De Nil, L. F., Sasisekaran, J., Van Lieshout, P. H., and Sandor, P. (2005). Speech disfluencies in individuals with Tourette syndrome. *Journal of Psychosomatic Research*, 58(1) :97–102.

Deshmukh, N., Ganapathiraju, A., Gleeson, A., Hamaker, J., and Picone, J. (1998). Resegmentation of SWITCHBOARD. In *The 5th International Conference on Spoken Language Processing*, Sydney, Australia.

Du Bois, J. W. (2007). The stance triangle. In Robert Englebretson, editor, *Stancetaking in*

Discourse : Subjectivity, evaluation, interaction, pages 139–182. John Benjamins Publishing Company, Amsterdam.

Duez, D. (1982). Silent and non-silent pauses in three speech styles. *Language and Speech*, 25(1) :11–28.

Englebretson, R. (2007). Stancetaking in discourse : An introduction. In Englebretson, R., editor, *Stancetaking in Discourse : Subjectivity, evaluation, interaction*, pages 1–25. John Benjamins Publishing Company, Amsterdam.

Ferguson, J., Durrett, G., and Klein, D. (2015). Disfluency detection with a semi-markov model and prosodic features. In *Proc. NAACL HLT*.

Fetterolf, F. and Marceau, M. (2013). A case of bupropion-induced stuttering.

Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of memory and language*, (34) :709–738.

Fox Tree, J. E. (1997). Pronouncing "the" as "thee" to signal problems in speaking. *Cognition*, 62(2) :151–167.

Fox Tree, J. E. (2001). Listeners' uses of um and uh in speech comprehension. *Memory & cognition*.

Freeman, V. (2014). Hyperarticulation as a signal of stance. *Journal of Phonetics*, 45(1) :1–11.

Freeman, V. (2015). *The Phonetics of Stance-taking*. PhD thesis, University of Washington, Seattle.

Freeman, V., Chan, J., Levow, G., and Wright, R. (2014a). ATAROS Technical Report 1 : Corpus collection and initial task validation. Technical report.

Freeman, V., Chan, J., Levow, G.-A., Wright, R., Ostendorf, M., and Zayats, V. (2014b). Manipulating stance and involvement using collabo- rative tasks : An exploratory comparison. In *Proceedings of Interspeech 2014*.

Freeman, V., Levow, G.-A., Wright, R., and Ostendorf, M. (2015a). Investigating the role of 'yeah' in stance-dense conversation. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech 2015)*, pages 2–6, Dresden.

Freeman, V., Wright, R., and Levow, G.-A. (2015b). The Prosody of negative 'yeah'. In *89th Annual Meeting of the Linguistics Society of America (LSA)*, Portland, OR.

Gabrea, M. and O'Shaughnessy, D. (2000). DETECTION OF FILLED PAUSES IN SPONTANEOUS CONVERSATIONAL SPEECH. *INTERSPEECH*, pages 678–681.

Georgila, K. (2009). Using Integer Linear Programming for Detecting Speech Disfluencies. In *NAACL HLT*, pages 109–112.

Giraud, A.-L., Neumann, K., Bachoud-Levi, A.-C., von Gudenberg, A. W., Euler, H. A., Lanfermann, H., and Preibisch, C. (2008). Severity of dysfluency correlates with basal ganglia activity in persistent developmental stuttering. *Brain and Language*, 104(2) :190–199.

Goberman, A. M., Blomgren, M., and Metzger, E. (2010). Characteristics of speech disfluency in Parkinson disease. *Journal of Neurolinguistics*, 23(5) :470–478.

Godfrey, J. and Holliman, E. (1993). Switchboard-1 Release 2 LDC97S62 [Corpus]. Technical report, Philadelphia : Linguistic Data Consortium.

Godfrey, J., Holliman, E., and McDaniel, J. (1992). SWITCHBOARD : telephone speech corpus for research and development. In *Proceedings of ICASSP-92 : 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1. IEEE.

Gorman, K., Olson, L., Hill, A. P., Lunsford, R., Heeman, P. A., and van Santen, J. P. H. (2016). Uh and um in children with autism spectrum disorders or language impairment. *Autism Research*, 9(8) :854–865.

Haddington, P. (2004). Stance Taking in News Interviews. *SKY Journal of Linguistics*, 17 :101–142.

Hamaker, J., Zeng, Y., and Picone, J. (1998). Rules and Guidelines for Transcription and Segmentation of the SWITCHBOARD Large Vocabulary Conversational Speech Recognition Corpus General Instructions for SWITCHBOARD Transcriptions. Technical report, Mississippi State University.

Hassan, H., Schwartz, L., Hakkani-Tür, D., and Tur, G. (2014). Segmentation and Disfluency Removal for Conversational Speech Translation. *INTERSPEECH*, pages 318–322.

Honal, M. and Schultz, T. (2005). Automatic Disfluency Removal on Recognized Spontaneous Speech - Rapid Adaptation to Speaker Dependent Disfluencies. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages 969–972. IEEE.

Hunston, S. and Thompson, G. (2000). *Evaluation in text : authorial stance and the construction of discourse*. Oxford University Press.

Irvine, C. A., Eigsti, I. M., and Fein, D. A. (2015). Uh, Um, and Autism : Filler Disfluencies as Pragmatic Markers in Adolescents with Optimal Outcomes from Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, pages 1–10.

Kadri, M., Balasubramanian, V., and Max, L. (2011). Loci of disfluency in acquired neurogenic versus persistent developmental stuttering. In *ASHA*, San Diego, CA.

Kidd, C., White, K. S., and Aslin, R. N. (2011). Toddlers use speech disfluencies to predict speakers' referential intentions. *Developmental Science*, 14(4) :925–934.

Kjellmer, G. (2003). Hesitation. In Defence of ER and ERM. *English Studies*, 84(2) :170–198.

Kumin, L. (1994). Intelligibility of Speech in Children with down Syndrome in Natural Settings : Parents' Perspective. *Perceptual and Motor Skills*, 78(1) :307–313.

Laserna, C. M., Seih, Y.-T., and Pennebaker, J. W. (2014). Um . . . Who Like Says You Know : Filler Word Use as a Function of Age, Gender, and Personality. *Journal of Language and Social Psychology*, 33(3) :328–338.

Lease, M., Johnson, M., and Charniak, E. (2006). Recognizing Disfluencies in Conversa-

tional Speech. *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, 14(5) :1566–1573.

Levow, G.-A., Freeman, V., Hrynkevich, A., Ostendorf, M., Wright, R., Chan, J., Luan, Y., and Tran, T. (2014). Recognition of stance strength and polarity in spontaneous speech. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 236–241, South Lake Tahoe, NV. IEEE.

Lickley, R. J. (1994). *Detecting Disuency in Spontaneous Speech*. PhD thesis, University of Edinburgh.

Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., and Harper, M. (2006a). Enriching speech recognition with automatic detection of sentence boundraries and disfluencies. *IEEE Transactions on audio, speech, and language processing*, 14 :1526–1540.

Liu, Y., Shriberg, E., Stolcke, A., Member, S., Hillard, D., Member, S., Ostendorf, M., and Harper, M. (2006b). Automatic Detection of Sentence Boundaries and Disfluencies. 14(5) :1526–1540.

Lutz, K. C. and Mallard, A. (1986). Disfluencies and rate of speech in young adult nonstutterers. *Journal of Fluency Disorders*, 11(4) :307–316.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of English : the penn treebank. *Computational Linguistics*, 19(2) :313–330.

Mareüil, P. B. D., Habert, B., Bénard, F., Adda-decker, M., Barras, C., Adda, G., and Paroubek, P. (2005). A quantitative study of disfluencies in French broadcast interviews. In *Disfluency in Spontaneous Speech*, number September, pages 27–32, Aix-en-Provence, France.

Marin, A. and Ostendorf, M. (2014). Domain adaptation for parsing in automatic speech recognition.

Meteer, M. and Taylor, A. (1995). Dysfluency Annotation Stylebook for the Switchboard Corpus. Technical report, University of Pennsylvania, Philadelphia, PA.

Moniz, H., Batista, F., Mata, A. I., and Trancoso, I. (2014). Speaking style effects in the production of disfluencies. *Speech Communication*, 65 :20–35.

Murray, G. and Carenini, G. (2009). Detecting subjectivity in multiparty speech. In *INTERSPEECH*, pages 2007–2010.

Nespor, M. and Vogel, I. (1986). *Prosodic phonology*. Foris, Dordrecht, Holland ; Riverton, N.J., U.S.A.

Noeth, E., Niemann, H., Haderlein, T., Decher, M., Eysholdt, U., Rosanowski, F., and Wittenberg, T. (2000). Automatic Stuttering Recognition using Hidden Markov Models. In *Conference on Spoken Language Processing*, volume 4, pages 65–68, Beijing, China.

Norrick, N. R. (2015). Interjections. In Aijmer, K. and Rühlemann, C., editors, *Corpus Pragmatics : A Handbook*, pages 249–275. Cambridge University Press, Cambridge.

O'Connell, D. C. and Kowal, S. (2005). Uh and um revisited : Are they interjections for signaling delay ? *Journal of Psycholinguistic Research*, 34(6) :555–576.

Ostendorf, M. and Hahn, S. (2013). A Sequential Repetition Model for Improved Disfluency Detection. *INTERSPEECH*, pages 2624–2628.

Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2) :1–135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up ? In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*, volume 10, pages 79–86, Morristown, NJ, USA. Association for Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn : Machine Learning in {P}ython. *Journal of Machine Learning Research*, 12 :2825–2830.

Preus, A. (1972). Stuttering in Down's Syndrome. *Scandinavian Journal of Educational Research*, 16(1) :89–104.

Quirk, R. (1985). *A Comprehensive grammar of the English language*. Longman.

R Core Team (2013). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Scaler Scott, K., Grossman, H. L., Abendroth, K. J., Tetnowski, J. A., and Damico, J. S. (2006). Asperger Syndrome and Attention Deficit Disorder : Clinical Disfluency Analysis. In *Proceedings of the 5th World Congress on Fluency Disorders*, Dublin, Ireland.

Schachter, S., Christenfeld, N., Ravina, B., and Bilous, F. (1991). Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology*, 60(3) :362.

Scheibman, J. (2002). *Point of View and Grammar*, volume 11 of *Studies in Discourse and Grammar*. John Benjamins Publishing Company, Amsterdam.

Scheibman, J. (2007). Subjective and intersubjective uses of generalizations in English conversations. In Englebretson, R., editor, *Stancetaking in Discourse : Subjectivity, evaluation, interaction*, pages 111–138. John Benjamins Publishing Company, Amsterdam.

Shriberg, E. (1994). *Preliminaries to a theory of speech disfluencies*. PhD thesis, University of California, Berkeley.

Shriberg, E. (1995). Acoustic Properties of Disfluent Repetitions. *Proceedings of International Conference on Phonetic Sciences (ICPhS)*, 4 :384–387.

Shriberg, E. (1996). Disfluencies in SWITCHBOARD. In *Proceedings of the International Conference on Spoken Language Processing*, pages 11–14, Philadelphia, PA.

Shriberg, E. (1999). Phonetic Consequences of Speech Disfluency. *International Congress of Phonetic Sciences*, pages 619–622.

Shriberg, E. (2001). To 'errrr' is human : ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31.

Shriberg, E. and Lickley, R. J. (1993). Intonation of Clause-Internal Filled Pauses. *Phonetica*, 50 :172–179.

Shriberg, E., Stolcke, A., and Baron, D. (2001a). Observations on Overlap : Findings and Implications for Automatic Processing of Multi-Party Conversation. In *INTERSPEECH*, pages 1359–1362.

Shriberg, L. D., Paul, R., McSweeny, J. L., Klin, A., Cohen, D. J., and Volkmar, F. R. (2001b). Speech and Prosody Characteristics of Adolescents and Adults With High-Functioning Autism and Asperger Syndrome. *Journal of Speech, Language, and Hearing Research*, 44(5) :1097–1115.

Sisskin, V. (2006). Speech Disfluency in Asperger's Syndrome : Two Cases of Interest. *SIG 4 Perspectives on Fluency and Fluency Disorders*, 16(2) :12–14.

Siu, M., Ostendorf, M., and Member, S. (2000). Variable N-Grams and Extensions for Conversational Speech Language Modeling. *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*, 8(1) :63–75.

Siu, M.-H. and Ostendorf, M. (1996). Modeling disfluencies in conversational speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 1, pages 386–389. IEEE.

Snover, M., Dorr, B., and Schwartz, R. (2004). A Lexically-Driven Algorithm for Disfluency Detection. *Hlt/Naacl*, pages 157–160.

Somasundaran, S. and Wiebe, J. (2009). Recognizing Stances in Online Debates. In *Proceedings of ACL 2009 : Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 226–234.

Somasundaran, S., Wiebe, J., Hoffmann, P., and Litman, D. (2006). Manual annotation of opinion categories in meetings. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, pages 54–61. Association for Computational Linguistics.

Starkweather, C. W. (1987). *Fluency and stuttering*. Prentice-Hall, englewood edition.

Stolcke, A. and Shriberg, E. (1996). Statistical language modeling for speech disfluencies. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 405–408. IEEE.

Stouten, F., Duchateau, J., Martens, J.-P., and Wambacq, P. (2006). Coping with disfluencies in spontaneous speech recognition : Acoustic detection and linguistic context manipulation. *Speech communication*, 48(1590–1606).

Strassel, S. (2004). Simple Metadata Annotation Specification V6.2. In *Linguistic Data Consortium*.

Sutton, C. and McCallum, A. (2012). An Introduction to Conditional Random Fields. *Foundations and Trends® in Machine Learning*, 4(4) :267–373.

Svartvik, J., editor (1990). *The London Corpus of Spoken English : Description and Research*. Lund University Press, Lund.

Swerts, M., Wichmann, A., and Beun, R.-J. (1998). Filled Pauses as Markers of Discourse Structure. *Journal of Pragmatics*, 30(4) :485–496.

Tottie, G. (2014). On the use of uh and um in American English. *Functions of Language*, 21(1) :6–29.

Tottie, G. (2015a). Turn management and the fillers uh and um. In *Corpus Pragmatics : A Handbook*, chapter 14, pages 381–407. Cambridge.

Tottie, G. (2015b). Uh and Um in British and American English : Are They Words ? Evidence from Co-occurrence with Pauses. In *Linguistic Variation : Confronting Fact and Theory*, pages 38–55.

Wang, W., Arora, R., Livescu, K., Bilmes, J., and Edu, B. W. (2015). On Deep Multi-View Representation Learning. In *32nd International Conference on Machine Learning, ICML*, pages 1083–1092.

Wang, W., Stolcke, A., Yuan, J., and Liberman, M. (2013). A Cross-language Study on Automatic Speech Disfluency Detection. In *NAACL-HLT*, number Dm, pages 703–708.

Watanabe, M., Hirose, K., Den, Y., and Minematsu, N. (2008). Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech Communication*, 50(2) :81–94.

Wilson, T. and Wiebe, J. (2005). Annotating attributions and private states. *Proceedings of the Workshop on Frontiers in Corpus*.

Wingate, M. E. (1964). A Standard Definition of Stuttering. *Journal of Speech and Hearing Disorders*, 29(4) :484–489.

Yuan, J. and Liberman, M. (2008). Speaker identification on the SCOTUS corpus. In *Journal of the Acoustical Society of America*.

Zayats, V., Ostendorf, M., and Hajishirzi, H. (2014). Multi-Domain Disfluency and Repair Detection. In *INTERSPEECH*, number September, pages 2907–2911.

Zayats, V., Ostendorf, M., and Hajishirzi, H. (2015). Unediting : Detecting Disfluencies Without Careful Transcripts. In *NAACL*, pages 1410–1415.

Zhang, L., Gao, Y., Hong, C., Feng, Y., Zhu, J., and Cai, D. (2014). Feature Correlation Hypergraph : Exploiting High-order Potentials for Multimodal Recognition. *IEEE Transactions on Cybernetics*, 44(8) :1408–1419.

Zwarts, S. and Johnson, M. (2011). The impact of language models and loss functions on repair disfluency detection. In *ACL*, pages 703–711.

Zwarts, S., Johnson, M., and Dale, R. (2010). Detecting speech repairs incrementally using a noisy channel approach. In *Coling*, pages 1371–1378, Beijing.

Appendix A

WORD CATEGORY : LIST OF *FUNCTION* AND *OTHER* WORDS

Tables A.1 and A.2 list the *other* and *function* words used in Chapter 7, detailed in section 7.2.4. The category *other* listed in Table A.1 was created to deal with words like backchannels, filled pauses, discourse markers, or interjections. These words have a high token frequency, similar to function words. Since they do not behave quite like function words, or like lexical words, I created the category *other* to account for these words in a separate category. The list of function words (see Table A.2) is a combination of the NLTK stopword list (Bird et al., 2009) and words manually added to match the tokenization used in the data processing, such as *na* from *wanna* or *gonna*, the negation *nt*, and contractions containing apostrophes like *'s, n't*, or *'re*.

Table A.1: List of *other* words

um	huh-uh	uh-huh	nah	hm
uh	hum	yeah	yep	eh
um-hum	hum-um	yes	oh	ooh
huh	uh-hum	nope	ah	

TABLE A.2: List of *function* words

i	their	doing	below	more	n't
me	theirs	a	to	most	're
my	themselves	an	from	other	'm
myself	what	the	up	some	've
we	which	and	down	such	'll
our	who	but	in	no	'd
ours	whom	if	out	nor	'
ourselves	this	or	on	not	'cause
you	that	because	off	only	'em
your	these	as	over	own	don't
yours	those	until	under	same	that's
yourself	am	while	again	so	they're,
yourselves	is	of	further	than	it's
he	are	at	then	too	isn't
him	was	by	once	very	aren't
his	were	for	here	s	didn't
himself	be	with	there	t	you'll
she	been	about	when	can	doesn't
her	being	against	where	will	i'm
hers	have	between	why	just	what's
herself	has	into	how	don	hadn't
it	had	through	all	should	can't
its	having	during	any	now	haven't
itself	do	before	both	nt	you're
they	does	after	each	na	
them	did	above	few	's	

Appendix B

TOKEN FREQUENCY OF MONOSYLLABIC WORDS WITH THE VOWEL /ʌ/ IN ATAROS

This appendix presents in Table B.1 the list of monosyllabic words with the vowel /ʌ/ (AH1), and their token frequency in the ATAROS corpus. This list is used in Chapter 9 to compare the vowels in *um* and *uh* to other monosyllabic words, labeled with the same vowel.

This list was compiled by using a Praat script (Boersma and Weenink, 2015) to identify the label AH1 in the vowel tier, and by extracting the corresponding word in the word tier (see section 5.3.4 for more information). The frequency was computed by adding the number of times a word is seen using R (R Core Team, 2013). Finally, I manually annotated the list for monosyllabic words, and only kept the monosyllabic ones.

This list shows that *um* and *uh* have the highest token frequency in the corpus, and that the token frequency of other monosyllabic words has a lot of variability.

TABLE B.1: Token frequency of monosyllabic words with the vowel /ʌ/ in ATAROS

Token	Freq	Token	Freq	Token	Freq	Token	Freq
UM	569	WHAT'S	18	JUMP	3	BUNS	1
UH	418	RUN	16	RUNS	3	CLU	1
BUT	360	US	16	SON	3	CUP	1
WHAT	300	JUST	12	SUCH	3	DUB	1
ONE	287	W	11	TRUCK	3	DUCKS	1
THE	264	FUND	10	BU	2	DUH	1
CUT	196	ONCE	10	BULBS	2	DUMP	1
UP	181	ONE'S	10	FRONT	2	DUN	1
STUFF	176	COMES	9	NONE	2	FLOOD	1
OF	149	FUN	9	PUB	2	FRONT'S	1
CUZ	145	TOUGH	9	SHUT	2	GLOVES	1
SOME	105	MUST	8	STA	2	HUNT	1
BUS	66	LUNCH	6	STUFF'S	2	LOVE	1
DONE	61	STUCK	6	STUFFS	2	NUTS	1
MUCH	59	CUTS	5	SUCKS	2	PLUG	1
CUPS	51	LUMP	5	U	2	PLUS	1
DUCT	40	WHA	5	UGH	2	PULP	1
HUH	37	BUNCH	4	YUM	2	PUNCH	1
JUGS	37	JUG	4	BLUNT	1	PUP	1
CLUB	35	JUNK	4	BUCK	1	SHOVE	1
ONES	22	UN	4	BUCKS	1	TON	1
WAS	21	WHAT'D	4	BUG	1	TOUCH	1
COME	20	CLUBS	3	BUGS	1	UNL	1
DOES	19	DRUG	3	BULK	1	UPS	1
FROM	18	FUNDS	3	BUN	1		

Appendix C

TOP 200 MOST FREQUENT WORDS IN ATAROS

This appendix presents in Table C.1 the list of 200 most frequent words in the ATAROS corpus and their frequency, in decreasing order. This list of words in used in Chapter 10 to optimize runtime, and to reduce the number of lexical features used in the classification experiment.

This list was compiled by using a Praat script (Boersma and Weenink, 2015) to extract the spurts in the transcription tiers (see section 5.3.4 for more information). The frequency was computed by creating a frequency dictionary in python, and by extracting the 200 words with the highest frequency.

TABLE C.1: Top 200 most frequent words in ATAROS in decreasing frequency order

Token	Freq	Token	Freq	Token	Freq	Token	Freq
yeah	2354	were	263	paper	106	bus	66
i	2005	on	252	down	106	cant	66
the	1822	good	251	at	103	power	65
that	1222	then	215	mm	102	definitely	65
we	1133	no	209	people	102	my	65
and	1105	important	202	out	101	large	64
of	917	probably	202	id	100	much	64
like	909	im	201	rid	99	sex	64
it	904	next	199	other	98	either	63
okay	882	cut	199	education	98	done	63
to	792	sure	190	stops	98	seems	63

a	786	mmhm	188	sounds	97	make	61
so	777	get	188	thing	96	else	61
you	675	see	187	boating	96	station	61
think	611	should	184	juice	95	near	60
thats	607	up	184	something	94	youre	60
um	593	keep	184	kinda	92	sense	60
is	547	gonna	183	football	90	access	60
dont	527	stuff	182	two	90	water	59
have	512	how	178	kind	86	want	59
well	483	really	177	dunno	86	pretty	59
here	481	public	177	because	85	had	59
oh	467	alright	176	me	84	did	59
in	460	those	171	your	83	cords	57
maybe	449	mean	159	true	83	aisle	57
with	442	yes	156	yep	83	same	57
uh	434	more	154	kay	82	ed	57
be	425	all	153	as	82	already	57
its	424	cuz	148	control	81	bout	57
know	416	was	147	from	80	area	56
would	403	hm	145	now	78	also	56
do	395	things	143	take	77	three	56
go	392	need	141	heavy	77	clothing	55
for	384	them	140	sticks	75	cups	55
or	382	food	137	any	75	weed	55
but	374	got	136	wanna	74	little	55
this	366	say	134	hospital	73	community	55
right	345	where	132	services	73	somewhere	54

put	345	by	128	feel	72	acting	54
there	343	too	127	taxi	72	might	53
could	329	about	126	small	71	cake	53
what	325	books	126	cream	70	knives	53
are	325	theyre	126	sugar	70	bars	53
if	314	guess	126	fishing	70	section	53
they	306	over	122	an	69	classes	53
just	295	theres	117	why	68	thinking	52
one	291	supplies	112	licenses	68	makes	52
can	290	these	109	box	67	soccer	52
not	270	some	108	news	67	money	52
lets	267	our	107	which	67	towing	52

Appendix D

MAIN PYTHON SCRIPTS USED IN CHAPTER 10

This appendix contains the four main scripts used in Chapter 10. I wrote Listings D.1 and D.3 to collect and format the features used in the two classification experiments. I wrote Listings D.2 and D.4 to preprocess the features, train and test the data, and to collect the scores.

Listing D.1 is used in section 10.3 to create and format the data with lexical and spurt information.

Listing D.1: Python script to extract the data for section 10.3

```
import csv
import os
import spurt_clean # script I made to clean text data
import operator
from collections import OrderedDict

# output the list into a csv format
def to_write(writer, list):
    for row in list:
        writer.writerow(row)

# returns a binary stance distinction
def get_bin(strength4):
    if strength4 == 2.0 or strength4 == 3.0:
        return 1.0
    else:
        return strength4

```

```
19 # returns a three-way stance distinction
20 def get_strength3(strength4):
21     if strength4 == 3.0:
22         return 2.0
23     else:
24         return strength4
25 
26 # returns the value of stance strength: None, 0, 1, 2 or 3
27 # and stance polarity
28 # 0 for negative, 1 for neutral and 2 for positive
29 def get_stance(stance):
30     if '0' in stance:
31         strength = 0.0
32     elif '1' in stance:
33         strength = 1.0
34     elif '2' in stance:
35         strength = 2.0
36     elif '3' in stance:
37         strength = 3.0
38     else:
39         strength = 4.0
40     if '-' in stance:
41         polarity = -1.0
42     elif '+' in stance:
43         polarity = 1.0
44     else:
45         polarity = 0.0
46     return strength, polarity
47 
48 # returns a clean spurt with no UM or UH
49 def remove_uhm(spurt):
50     tokenized = spurt.lower().strip().split()
51     scrappy = list()
```

```
52      for item in tokenized:
53          if item != 'um' and item != 'uh':
54              scrappy.append(item)
55      return ' '.join(scrappy)
56
57  # return 0 for Inventory task '3I'
58  # and 1 for Budget task '6B'
59  def get_task(task):
60      if task == '6B':
61          return 1.0
62      else:
63          return 0.0
64
65  # returns 0 for male, 1 for female
66  def get_gender(speaker):
67      if speaker.startswith('NWF'):
68          return 1.0
69      else:
70          return 0.0
71
72  # returns the spurts with only words
73  # that are among the 200 most frequent
74  def transform_spurt(clean, top200):
75      new_clean = list()
76      splitted = clean.lower().strip().split()
77      for item in splitted:
78          if item in top200:
79              new_clean.append(item)
80      transformed = ' '.join(new_clean)
81      return transformed
82
83  # returns a clean spurt
84  def to_clean(raw):
```

```
85      clean = spurt_clean.remove_parens(raw)
86      clean = spurt_clean.remove_truncated(clean)
87      clean = spurt_clean.strip_punct(clean)
88      clean = spurt_clean.makes_one_space(clean)
89      return clean
90
91  # returns a sorted dictionary
92  # with top 200 most frequent words
93  def get_top200(myDict):
94      top200_dict = dict()
95      sorted_d = sorted(myDict.items(), key=lambda kv: kv[1], reverse=True)
96      for k,v in sorted_d[:200]:
97          top200_dict[k] = v
98      return top200_dict
99
100 # returns a frequency dictionary for all spurts
101 def get_freq(data):
102     freq_dict = dict()
103     for line in data:
104         line = line.split('\t')
105         spurt = line[2]
106         clean = to_clean(spurt)
107         tokenized = clean.strip().lower().split()
108         if len(clean) > 0:
109             for item in tokenized:
110                 if item in freq_dict:
111                     freq_dict[item] += 1
112                 else:
113                     freq_dict[item] = 1
114     return freq_dict
115
116 def process(data, writer):
117     freq_dict = get_freq(data) # get the frequency dictionary
```

```
118     # get the most frequent 200 words
119     # from the frequency dictionary
120     top200 = get_top200(freq_dict)
121     newlist = list()
122     for i, line in enumerate(data):
123         temp = list()
124         if i > 0: # to skip data header
125             splitted = line.strip().split('\t')
126             # ignore lines with missing data
127             if len(splitted) > 5:
128                 speaker = splitted[0]
129                 task = splitted[1]
130                 raw = splitted[2]
131                 dur = splitted[3]
132                 token = splitted[4]
133                 coarse = splitted[5]
134                 gender = get_gender(speaker)
135                 strength4, polarity = get_stance(coarse)
136                 new_task = get_task(task)
137                 strength3 = get_strength3(strength4)
138                 bin = get_bin(strength4)
139                 clean = to_clean(raw) # clean text from spurts
140                 # create a copy of spurts with only
141                 # the most frequent words
142                 transformed = transform_spurt(clean, top200)
143                 # avoid undetermined stance and
144                 # empty spurts (some spurts only have comments)
145                 if strength4 != 4.0 and len(clean) > 0:
146                     scrapped = remove_uhm(clean) # filter UM and UH
147                     if len(scrapped) > 0 and len(transformed) > 0:
148                         temp.extend([scrapped, transformed, gender, new_task,
149                                      dur, strength4, polarity, bin, strength3])
150                         newlist.append(temp) # create new data
```

```
151     to_write(writer, newlist)
152 
153 # returns a list of lines from the data
154 def get_data(data):
155     return [line.strip() for line in open(data, 'r')]
156 
157 # returns a writer object to output to the csv
158 # with a header
159 def make_csv(dir):
160     to_write = open(dir + os.sep + 'lexical_no_uhm.csv', 'w', newline='')
161     writer = csv.writer(to_write, delimiter=',')
162     writer.writerow(['clean', 'top200', 'gender', 'task', 'duration',
163     'strength4', 'polarity', 'bin', 'strength3'])
164     return writer
165 
166 def main():
167     spurts = '<inputData>'
168     out_path = '<outPutDirectory>'
169     writer = make_csv(out_path)
170     dat = get_data(spurts)
171     process(dat, writer)
172 main()
```

Listing D.2 is a sample code used in section 10.3 to preprocess the features, fit the classifier, and get the accuracy, precision, recall, and F1-scores over five folds of the data.

Listing D.2: Python script to train the classifier for section 10.3

```
1 
2 import pandas as pd
3 import numpy as np
4 from sklearn.svm import SVC
5 from sklearn import preprocessing
6 from sklearn.feature_extraction.text import CountVectorizer
7 from sklearn import metrics
```

```
8
9  # reads in the data as pandas
10 dat = pd.read_csv('<data>')
11
12 # get the target data
13 y_pol = dat.polarity.values
14 y_bin = dat.bin.values
15 y_str3 = dat.strength3.values
16
17 # scale and reshape numerical features
18 gen_scaled = preprocessing.scale(dat['gender'].values)
19 gen_scaled = gen_scaled.reshape(-1,1)
20 task_scaled = preprocessing.scale(dat['task'].values)
21 task_scaled = task_scaled.reshape((-1,1))
22 dur_scaled = preprocessing.scale(dat['duration'].values)
23 dur_scaled = dur_scaled.reshape(-1, 1)
24
25 # vectorize text features
26 count_vect = CountVectorizer()
27 bow = count_vect.fit_transform(dat['top200'].values)
28
29 # Setting the data
30 # use all features
31 X_all = np.hstack([bow.toarray(), gen_scaled, task_scaled, dur_scaled])
32 # use no lexical features
33 X_no_lex = np.hstack([gen_scaled, task_scaled, dur_scaled])
34
35 # initialize classifier with no weight:
36 clf = SVC(kernel='linear', gamma=0.001, C=1.0)
37
38 # initialize classifier with weights
39 wclf_1 = SVC(kernel='linear', class_weight="balanced")
40 wclf_2 = SVC(kernel='linear', class_weight={0:.60, 1:.40}, gamma=0.001, C=1.0)
```

```
41
42 # BINARY STANCE
43
44 # TESTING BINARY STANCE (y_bin) WITH ALL FEATURES (X_all)
45 # divide data into 5 folds to fit and predict
46 X_folds = np.array_split(X_all, 5)
47 y_folds = np.array_split(y_bin, 5)
48 scores = list()
49 for k in range(5):
50     X_train = list(X_folds)
51     X_test = X_train.pop(k)
52     X_train = np.concatenate(X_train)
53     y_train = list(y_folds)
54     y_test = y_train.pop(k)
55     y_train = np.concatenate(y_train)
56     scores.append(wclf_2.fit(X_train,y_train).score(X_test,y_test))
57     predictions = wclf_2.predict(X_test)
58     # print accuracy
59     print(round(np.mean(predictions == y_test) * 100, 2))
60     # print precision, recall, F1
61     print(metrics.classification_report(y_test, predictions, digits=3))
62     # print confusion matrix
63     print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
    Predicted'], margins=True))
64 # print list of k accuracy scores and mean of scores
65 scores_mean = sum(scores)/float(len(scores))
66 print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')
67
68 # TESTING BINARY STANCE (y_bin) WITH ONLY LEXICAL FEATURES (bow)
69 # divide data into 5 folds to fit and predict
70 X_folds = np.array_split(bow.toarray(), 5)
71 y_folds = np.array_split(y_bin, 5)
72 scores = list()
```

```
for k in range(5):
    X_train = list(X_folds)
    X_test = X_train.pop(k)
    X_train = np.concatenate(X_train)
    y_train = list(y_folds)
    y_test = y_train.pop(k)
    y_train = np.concatenate(y_train)
    scores.append(wclf_2.fit(X_train,y_train).score(X_test,y_test))
    predictions = wclf_2.predict(X_test)
    # print accuracy
    print(round(np.mean(predictions == y_test) * 100, 2))
    # print precision, recall, F1
    print(metrics.classification_report(y_test, predictions, digits=3))
    # print confusion matrix
    print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
   Predicted'], margins=True))
# print list of k accuracy scores and mean of scores
scores_mean = sum(scores)/float(len(scores))
print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')

# TESTING BINARY STANCE (y_bin) WITH ONLY LEXICAL FEATURES (X_no_lex)
# divide data into 5 folds to fit and predict
X_folds = np.array_split(X_no_lex, 5)
y_folds = np.array_split(y_bin, 5)
scores = list()
for k in range(5):
    X_train = list(X_folds)
    X_test = X_train.pop(k)
    X_train = np.concatenate(X_train)
    y_train = list(y_folds)
    y_test = y_train.pop(k)
    y_train = np.concatenate(y_train)
    scores.append(wclf_2.fit(X_train,y_train).score(X_test,y_test))
```

```
105     predictions = wclf_2.predict(X_test)
106     # print accuracy
107     print(round(np.mean(predictions == y_test) * 100, 2))
108     # print precision, recall, F1
109     print(metrics.classification_report(y_test, predictions, digits=3))
110     # print confusion matrix
111     print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
        Predicted'], margins=True))
112 # print list of k accuracy scores and mean of scores
113 scores_mean = sum(scores)/float(len(scores))
114 print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')
115
116
117 # 3 WAY STANCE STRENGTH
118
119 wclf_3 = SVC(kernel='linear', class_weight={0:.4, 1:.2, 2:.4}, gamma=0.001, C
        =1.0)
120
121 # TESTING 3 WAY STANCE (y_str3) WITH ALL FEATURES (X_all)
122 # divide data into 5 folds to fit and predict
123 X_folds = np.array_split(X_all, 5)
124 y_folds = np.array_split(y_str3, 5)
125 scores = list()
126 for k in range(5):
127     X_train = list(X_folds)
128     X_test = X_train.pop(k)
129     X_train = np.concatenate(X_train)
130     y_train = list(y_folds)
131     y_test = y_train.pop(k)
132     y_train = np.concatenate(y_train)
133     scores.append(wclf_3.fit(X_train,y_train).score(X_test,y_test))
134     predictions = wclf_3.predict(X_test)
135     # print accuracy
```

```
136     print(round(np.mean(predictions == y_test) * 100, 2))
137     # print precision, recall, F1
138     print(metrics.classification_report(y_test, predictions, digits=3))
139     # print confusion matrix
140     print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
        Predicted'], margins=True))
141 # print list of k accuracy scores and mean of scores
142 scores_mean = sum(scores)/float(len(scores))
143 print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')
144
145 # TESTING 3 WAY STANCE (y_str3) WITH ONLY LEXICAL FEATURES (bow)
146 # divide data into 5 folds to fit and predict
147 X_folds = np.array_split(bow.toarray(), 5)
148 y_folds = np.array_split(y_str3, 5)
149 scores = list()
150 for k in range(5):
151     X_train = list(X_folds)
152     X_test = X_train.pop(k)
153     X_train = np.concatenate(X_train)
154     y_train = list(y_folds)
155     y_test = y_train.pop(k)
156     y_train = np.concatenate(y_train)
157     scores.append(wclf_3.fit(X_train,y_train).score(X_test,y_test))
158     predictions = wclf_3.predict(X_test)
159     # print accuracy
160     print(round(np.mean(predictions == y_test) * 100, 2))
161     # print precision, recall, F1
162     print(metrics.classification_report(y_test, predictions, digits=3))
163     # print confusion matrix
164     print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
        Predicted'], margins=True))
165 # print list of k accuracy scores and mean of scores
166 scores_mean = sum(scores)/float(len(scores))
```

```
print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')

# TESTING 3 WAY STANCE (y_str3) WITH ONLY LEXICAL FEATURES (X_no_lex)
# divide data into 5 folds to fit and predict
X_folds = np.array_split(X_no_lex, 5)
y_folds = np.array_split(y_str3, 5)
scores = list()
for k in range(5):
    X_train = list(X_folds)
    X_test = X_train.pop(k)
    X_train = np.concatenate(X_train)
    y_train = list(y_folds)
    y_test = y_train.pop(k)
    y_train = np.concatenate(y_train)
    scores.append(wclf_3.fit(X_train,y_train).score(X_test,y_test))
    predictions = wclf_3.predict(X_test)
    # print accuracy
    print(round(np.mean(predictions == y_test) * 100, 2))
    # print precision, recall, F1
    print(metrics.classification_report(y_test, predictions, digits=3))
    # print confusion matrix
    print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
   Predicted'], margins=True))
# print list of k accuracy scores and mean of scores
scores_mean = sum(scores)/float(len(scores))
print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')

# STANCE POLARITY

wclf_4 = SVC(kernel='linear', class_weight={-1:.55, 0:.15, 1:.3}, gamma=0.001,
    C=1.0)

# TESTING STANCE POLARITY (y_pol) WITH ALL FEATURES (X_all)
```

```
198 # divide data into 5 folds to fit and predict
199 X_folds = np.array_split(X_all, 5)
200 y_folds = np.array_split(y_pol, 5)
201 scores = list()
202 for k in range(5):
203     X_train = list(X_folds)
204     X_test = X_train.pop(k)
205     X_train = np.concatenate(X_train)
206     y_train = list(y_folds)
207     y_test = y_train.pop(k)
208     y_train = np.concatenate(y_train)
209     scores.append(wclf_4.fit(X_train,y_train).score(X_test,y_test))
210     predictions = wclf_4.predict(X_test)
211     # print accuracy
212     print(round(np.mean(predictions == y_test) * 100, 2))
213     # print precision, recall, F1
214     print(metrics.classification_report(y_test, predictions, digits=3))
215     # print confusion matrix
216     print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
    Predicted'], margins=True))
217 # print list of k accuracy scores and mean of scores
218 scores_mean = sum(scores)/float(len(scores))
219 print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')
220
221 # TESTING STANCE POLARITY (y_pol) WITH ONLY LEXICAL FEATURES (bow)
222 # divide data into 5 folds to fit and predict
223 X_folds = np.array_split(bow.toarray(), 5)
224 y_folds = np.array_split(y_pol, 5)
225 scores = list()
226 for k in range(5):
227     X_train = list(X_folds)
228     X_test = X_train.pop(k)
229     X_train = np.concatenate(X_train)
```

```
230     y_train = list(y_folds)
231     y_test = y_train.pop(k)
232     y_train = np.concatenate(y_train)
233     scores.append(wclf_4.fit(X_train,y_train).score(X_test,y_test))
234     predictions = wclf_4.predict(X_test)
235     # print accuracy
236     print(round(np.mean(predictions == y_test) * 100, 2))
237     # print precision, recall, F1
238     print(metrics.classification_report(y_test, predictions, digits=3))
239     # print confusion matrix
240     print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
        Predicted'], margins=True))
241 # print list of k accuracy scores and mean of scores
242 scores_mean = sum(scores)/float(len(scores))
243 print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')
244
245 # TESTING STANCE POLARITY (y_pol) WITH ONLY LEXICAL FEATURES (X_no_lex)
246 # divide data into 5 folds to fit and predict
247 X_folds = np.array_split(X_no_lex, 5)
248 y_folds = np.array_split(y_pol, 5)
249 scores = list()
250 for k in range(5):
251     X_train = list(X_folds)
252     X_test = X_train.pop(k)
253     X_train = np.concatenate(X_train)
254     y_train = list(y_folds)
255     y_test = y_train.pop(k)
256     y_train = np.concatenate(y_train)
257     scores.append(wclf_4.fit(X_train,y_train).score(X_test,y_test))
258     predictions = wclf_4.predict(X_test)
259     # print accuracy
260     print(round(np.mean(predictions == y_test) * 100, 2))
261     # print precision, recall, F1
```

```
    print(metrics.classification_report(y_test, predictions, digits=3))
    # print confusion matrix
    print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
    Predicted'], margins=True))
# print list of k accuracy scores and mean of scores
scores_mean = sum(scores)/float(len(scores))
print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')
```

Listing D.3 is used to create the data for the classification experiment on acoustic and position information, used in section 10.4.

Listing D.3: Python script to extract the data for section 10.4

```
import csv
import re
import spurt_clean # script I made to clean text data

# output data to csv file
def write_out(dat, out):
    for row in dat:
        out.writerow(row)

# collect and organize the data to output
def to_format(dat):
    new_dat = list()
    for row in dat:
        temp = list()
        id = row[0]
        raw = row[1]
        clean = row[22]
        speaker = row[2]
        gender = row[3]
        task = row[4]
        word = row[6]
        interval = row[7]
```

```
23          dur = row[9]
24           f0 = row[10]
25          f1 = row[11]
26          f2 = row[12]
27          intensity = row[13]
28          strength4 = row[15]
29          strength3 = row[16]
30          bin = row[17]
31          pol = row[18]
32          index = row[19]
33          totUHM = row[20]
34          position = row[21]
35          freq = row[23]
36          mono = row[24]
37          temp.extend([id, raw, clean, speaker, gender, task, word, interval,
38                       dur, f0, f1, f2, intensity,
39                       strength4, strength3, bin, pol,
40                       index, totUHM, freq, mono,
41                       position[0], position[1], position[2], position[3]])
42          new_dat.append(temp)
43      return new_dat
44
45  # returns a vector of binary values for position
46  def get_position(splitted, index):
47      length = len(splitted)
48      if length == 1:
49          position = [1, 0, 0, 0]
50      elif index == length - 1:
51          position = [0, 0, 0, 1]
52      elif index == 0:
53          position = [0, 1, 0, 0]
54      else:
55          position = [0, 0, 1, 0]
```

```
56      return position
57
58  # get the token frequency
59  def get_mono_freq(data, syll_dat):
60      new_dat = list()
61      for row in data:
62          word = row[6]
63          for line in syll_dat:
64              token = line[0]
65              if token == word:
66                  row.extend([line[1], line[2]])
67                  new_dat.append(row)
68      return new_dat
69
70  # to know which words have been processed when
71  # they are in the same spurt
72  def get_index(splitSpurt, word, currentSet):
73      for i, token in enumerate(splitSpurt):
74          if token.strip() == word and i not in currentSet:
75              currentSet.add(i)
76              return i
77      return None
78
79  # returns the total number of UM or UH in a spurt
80  def get_tot_uhm(splitted):
81      tot = 0
82      for item in splitted:
83          if item == 'um' or item == 'uh':
84              tot += 1
85      return tot
86
87  # returns a clean spurt
88  def to_clean(spurt):
```

```
89      clean = spurt_clean.remove_parens(spurt)
90      clean = spurt_clean.strip_punct(clean)
91      clean = spurt_clean.makes_one_space(clean)
92      return clean
93
94  # returns a list of indexed data for the vowels in the spurt
95  def index_data(data):
96      newList = []
97      priorInterval = '-1'
98      currentSet = set()
99      check = 0
100     for dataList in data:
101         flag = False
102         if priorInterval != dataList[7]:
103             priorInterval = dataList[7]
104             currentSet = set()
105         clean = to_clean(dataList[1])
106         splitSpurt = clean.lower().split()
107         totalUhm = get_tot_uhm(splitSpurt)
108         word = dataList[6].lower().strip()
109         word = re.sub('\'', '', word)
110         index = get_index(splitSpurt, word, currentSet)
111
112         # 112 rows have a none index, which means I am
113         # not interested in the word, so I don't append it
114
115         if index != None:
116             position = get_position(splitSpurt, index)
117             dataList.extend([index, totalUhm, position, clean])
118             newList.append(dataList)
119     return newList
120
121 def get_data(dat, syll, out):
```

```
122     indexed_data = index_data(dat)
123     mono_freq_dat = get_mono_freq(indexed_data, syll)
124     formated_dat = to_format(mono_freq_dat)
125     write_out(formated_dat, out)
126 
127 # initiates the output file with header
128 # returns a csv writer object
129 def make_out(fileName):
130     to_write = open(fileName, 'w', newline='')
131     writer = csv.writer(to_write, delimiter=',')
132     writer.writerow(['id', 'raw', 'clean', 'speaker', 'gender', 'task',
133                      'word', 'interval',
134                      'duration', 'f0', 'f1', 'f2', 'intensity',
135                      'strength4', 'strength3', 'bin', 'pol',
136                      'index', 'totalUHM','freq', 'monosyll',
137                      'alone', 'initial', 'medial', 'final'])
138     return writer
139 
140 # returns a list of data from the csv file
141 def to_read(file):
142     reader = csv.reader(file)
143     return [line for line in reader]
144 
145 def master(data, syll, outFile):
146     dat = to_read(data)
147     mono_syll = to_read(syll)
148     writer = make_out(outFile)
149     get_data(dat, mono_syll, writer)
150 
151 def main():
152     data = open(<inputData>)
153     syllables = open(<syllable_data>)
154     output = '<outPutName>'
```

```
    master(data, syllables, output)
main()
```

Listing D.4 is a sample code used in section 10.4 to preprocess the features, fit the classifier, and get the accuracy, precision, recall, and F1-scores over five folds of the data.

Listing D.4: Python script to train the classifier for section 10.4

```

import pandas as pd
import numpy as np
from sklearn.svm import SVC
from sklearn import preprocessing
from sklearn import metrics

# reads in the data as pandas
dat = pd.read_csv('<data>')

# get the target data
y_pol = dat.pol.values
y_bin = dat.bin.values
y_str3 = dat.strength3.values

# scale and reshape numerical features
dur_scaled = preprocessing.scale(dat['duration'].values)
dur_scaled = dur_scaled.reshape(-1, 1)
f1_scaled = preprocessing.scale(dat['f1'].values)
f1_scaled = f1_scaled.reshape(-1, 1)
f2_scaled = preprocessing.scale(dat['f2'].values)
f2_scaled = f2_scaled.reshape(-1, 1)
intensity_scaled = preprocessing.scale(dat['intensity'].values)
intensity_scaled = intensity_scaled.reshape(-1, 1)
alone_scaled = preprocessing.scale(dat['alone'].values)
alone_scaled = alone_scaled.reshape(-1, 1)
initial_scaled = preprocessing.scale(dat['initial'].values)
```

```
initial_scaled = initial_scaled.reshape(-1, 1)
medial_scaled = preprocessing.scale(dat['medial'].values)
medial_scaled = medial_scaled.reshape(-1, 1)
final_scaled = preprocessing.scale(dat['final'].values)
final_scaled = final_scaled.reshape(-1, 1)
word_scaled = preprocessing.scale(dat['uhm_bin'].values)
word_scaled = word_scaled.reshape(-1, 1)
gen_scaled = preprocessing.scale(dat['gen_bin'].values)
gen_scaled = gen_scaled.reshape(-1,1)
task_scaled = preprocessing.scale(dat['task_bin'].values)
task_scaled = task_scaled.reshape((-1,1))

# use all features
X_all = np.hstack([dur_scaled,
                   f1_scaled, f2_scaled,
                   intensity_scaled, alone_scaled,
                   initial_scaled, medial_scaled,
                   final_scaled, word_scaled,
                   gen_scaled, task_scaled])

# use only acoustic features
X_acous = np.hstack([dur_scaled,
                   f1_scaled, f2_scaled,
                   intensity_scaled])

# use only position features
X_posi = np.hstack([alone_scaled,
                   initial_scaled, medial_scaled,
                   final_scaled])

# use no position features
X_no_posi = np.hstack([dur_scaled,
                   f1_scaled, f2_scaled,
```

```
61                      intensity_scaled, word_scaled,
62                      gen_scaled, task_scaled])
63
64 # use no alone features
65 X_no_alone = np.hstack([dur_scaled,
66                      f1_scaled, f2_scaled,
67                      intensity_scaled,
68                      initial_scaled, medial_scaled,
69                      final_scaled, word_scaled,
70                      gen_scaled, task_scaled])
71
72 # use no initial features
73 X_no_initial = np.hstack([dur_scaled,
74                      f1_scaled, f2_scaled,
75                      intensity_scaled, alone_scaled,
76                      medial_scaled,
77                      final_scaled, word_scaled,
78                      gen_scaled, task_scaled])
79
80 # use no medial features
81 X_no_medial = np.hstack([dur_scaled,
82                      f1_scaled, f2_scaled,
83                      intensity_scaled, alone_scaled,
84                      initial_scaled,
85                      final_scaled, word_scaled,
86                      gen_scaled, task_scaled])
87
88 # use no final features
89 X_no_final = np.hstack([dur_scaled,
90                      f1_scaled, f2_scaled,
91                      intensity_scaled, alone_scaled,
92                      initial_scaled, medial_scaled,
93                      word_scaled,
```

```
94                      gen_scaled, task_scaled])
95
96  # use no alone+initial features
97  X_no_al_ini = np.hstack([dur_scaled,
98                      f1_scaled, f2_scaled,
99                      intensity_scaled, medial_scaled,
100                     final_scaled, word_scaled,
101                     gen_scaled, task_scaled])
102
103 # use no medial+final features
104 X_no_med_fin = np.hstack([dur_scaled,
105                     f1_scaled, f2_scaled,
106                     intensity_scaled, alone_scaled,
107                     initial_scaled, word_scaled,
108                     gen_scaled, task_scaled])
109
110 # use no initial+medial features
111 X_no_ini_med = np.hstack([dur_scaled,
112                     f1_scaled, f2_scaled,
113                     intensity_scaled, alone_scaled,
114                     final_scaled,
115                     word_scaled,
116                     gen_scaled, task_scaled])
117
118 # use no alone+initial+medial features
119 X_no_al_ini_med = np.hstack([dur_scaled,
120                     f1_scaled, f2_scaled,
121                     intensity_scaled,
122                     final_scaled, word_scaled,
123                     gen_scaled, task_scaled])
124
125 # use no word label (i.e., um vs. uh)
126 X_no_word = np.hstack([dur_scaled,
```

```
127                     f1_scaled, f2_scaled,
128                     intensity_scaled, alone_scaled,
129                     initial_scaled, medial_scaled,
130                     final_scaled,
131                     gen_scaled, task_scaled])
132
133 # use no acoustic features
134 X_no_acoustics = np.hstack([alone_scaled,
135                             initial_scaled, medial_scaled,
136                             final_scaled, word_scaled,
137                             gen_scaled, task_scaled])
138
139 # use no duration features
140 X_no_dur = np.hstack([f1_scaled, f2_scaled,
141                     intensity_scaled, alone_scaled,
142                     initial_scaled, medial_scaled,
143                     final_scaled, word_scaled,
144                     gen_scaled, task_scaled])
145
146 # use fo f1 features
147 X_no_f1 = np.hstack([dur_scaled,
148                     f2_scaled,
149                     intensity_scaled, alone_scaled,
150                     initial_scaled, medial_scaled,
151                     final_scaled, word_scaled,
152                     gen_scaled, task_scaled])
153
154 # use no f2 features
155 X_no_f2 = np.hstack([dur_scaled,
156                     f1_scaled,
157                     intensity_scaled, alone_scaled,
158                     initial_scaled, medial_scaled,
159                     final_scaled, word_scaled,
```

```
                    gen_scaled, task_scaled])

# use no intensity features
X_no_itensity = np.hstack([dur_scaled,
                    f1_scaled, f2_scaled,
                    alone_scaled,
                    initial_scaled, medial_scaled,
                    final_scaled, word_scaled,
                    gen_scaled, task_scaled])

# use no gender features
X_no_gen = np.hstack([dur_scaled,
                    f1_scaled, f2_scaled,
                    intensity_scaled, alone_scaled,
                    initial_scaled, medial_scaled,
                    final_scaled, word_scaled,
                    task_scaled])

# use no task
X_no_task = np.hstack([dur_scaled,
                    f1_scaled, f2_scaled,
                    intensity_scaled, alone_scaled,
                    initial_scaled, medial_scaled,
                    final_scaled, word_scaled,
                    gen_scaled])

# initialize classifier
clf = SVC(kernel='linear', gamma=0.001, C=1.0)

# initialize classifier with weights
wclf_1 = SVC(kernel='linear', class_weight="balanced")
wclf_2 = SVC(kernel='linear', class_weight={0:.6, 1:.4}, gamma=0.001, C=1.0)

```

```
193 # BINARY STANCE
194
195 # TESTING BINARY STANCE (y_bin) WITH ALL FEATURES (X_all)
196 # divide data into 5 folds to fit and predict
197 X_folds = np.array_split(X_all, 5)
198 y_folds = np.array_split(y_bin, 5)
199 scores = list()
200 for k in range(5):
201     X_train = list(X_folds)
202     X_test = X_train.pop(k)
203     X_train = np.concatenate(X_train)
204     y_train = list(y_folds)
205     y_test = y_train.pop(k)
206     y_train = np.concatenate(y_train)
207     scores.append(wclf_2.fit(X_train, y_train).score(X_test, y_test))
208     predictions = wclf_2.predict(X_test)
209     # print accuracy
210     print(round(np.mean(predictions == y_test) * 100, 2))
211     # print precision, recall, F1
212     print(metrics.classification_report(y_test, predictions, digits=3))
213     # print confusion matrix
214     print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
    Predicted'], margins=True))
215 # print list of k accuracy scores and mean of scores
216 scores_mean = sum(scores)/float(len(scores))
217 print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')
218
219 # TESTING BINARY STANCE (y_bin) WITH NO POSITION (X_posi)
220 # divide data into 5 folds to fit and predict
221 X_folds = np.array_split(X_posi, 5)
222 y_folds = np.array_split(y_bin, 5)
223 scores = list()
224 for k in range(5):
```

```
225     X_train = list(X_folds)
226     X_test = X_train.pop(k)
227     X_train = np.concatenate(X_train)
228     y_train = list(y_folds)
229     y_test = y_train.pop(k)
230     y_train = np.concatenate(y_train)
231     scores.append(wclf_2.fit(X_train, y_train).score(X_test, y_test))
232     predictions = wclf_2.predict(X_test)
233     # print accuracy
234     print(round(np.mean(predictions == y_test) * 100, 2))
235     # print precision, recall, F1
236     print(metrics.classification_report(y_test, predictions, digits=3))
237     # print confusion matrix
238     print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
    Predicted'], margins=True))
239 # print list of k accuracy scores and mean of scores
240 scores_mean = sum(scores)/float(len(scores))
241 print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')
242
243 # TESTING BINARY STANCE (y_bin) WITH NO POSITION (X_acous)
244 # divide data into 5 folds to fit and predict
245 X_folds = np.array_split(X_acous, 5)
246 y_folds = np.array_split(y_bin, 5)
247 scores = list()
248 for k in range(5):
249     X_train = list(X_folds)
250     X_test = X_train.pop(k)
251     X_train = np.concatenate(X_train)
252     y_train = list(y_folds)
253     y_test = y_train.pop(k)
254     y_train = np.concatenate(y_train)
255     scores.append(wclf_2.fit(X_train, y_train).score(X_test, y_test))
256     predictions = wclf_2.predict(X_test)
```

```
257     # print accuracy
258     print(round(np.mean(predictions == y_test) * 100, 2))
259     # print precision, recall, F1
260     print(metrics.classification_report(y_test, predictions, digits=3))
261     # print confusion matrix
262     print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
        Predicted'], margins=True))
263 # print list of k accuracy scores and mean of scores
264 scores_mean = sum(scores)/float(len(scores))
265 print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')
266
267 # TESTING BINARY STANCE (y_bin) WITH NO POSITION (X_no_acoustics)
268 # divide data into 5 folds to fit and predict
269 X_folds = np.array_split(X_no_acoustics, 5)
270 y_folds = np.array_split(y_bin, 5)
271 scores = list()
272 for k in range(5):
273     X_train = list(X_folds)
274     X_test = X_train.pop(k)
275     X_train = np.concatenate(X_train)
276     y_train = list(y_folds)
277     y_test = y_train.pop(k)
278     y_train = np.concatenate(y_train)
279     scores.append(wclf_2.fit(X_train, y_train).score(X_test, y_test))
280     predictions = wclf_2.predict(X_test)
281     # print accuracy
282     print(round(np.mean(predictions == y_test) * 100, 2))
283     # print precision, recall, F1
284     print(metrics.classification_report(y_test, predictions, digits=3))
285     # print confusion matrix
286     print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
        Predicted'], margins=True))
287 # print list of k accuracy scores and mean of scores
```

```
scores_mean = sum(scores)/float(len(scores))
print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')

# TESTING BINARY STANCE (y_bin) WITH NO POSITION (X_no_posi)
# divide data into 5 folds to fit and predict
X_folds = np.array_split(X_no_posi, 5)
y_folds = np.array_split(y_bin, 5)
scores = list()
for k in range(5):
    X_train = list(X_folds)
    X_test = X_train.pop(k)
    X_train = np.concatenate(X_train)
    y_train = list(y_folds)
    y_test = y_train.pop(k)
    y_train = np.concatenate(y_train)
    scores.append(wclf_2.fit(X_train, y_train).score(X_test, y_test))
    predictions = wclf_2.predict(X_test)
    # print accuracy
    print(round(np.mean(predictions == y_test) * 100, 2))
    # print precision, recall, F1
    print(metrics.classification_report(y_test, predictions, digits=3))
    # print confusion matrix
    print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
   Predicted'], margins=True))
# print list of k accuracy scores and mean of scores
scores_mean = sum(scores)/float(len(scores))
print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')

# 3 WAY STANCE STRENGTH

wclf_3 = SVC(kernel='linear', class_weight={0:.26, 1:.37, 2:.37}, gamma=0.001,
    C=1.0)

```

```
319 # TESTING 3 WAY STANCE (y_str3) WITH ALL FEATURES (X_all)
320 # divide data into 5 folds to fit and predict
321 X_folds = np.array_split(X_all, 5)
322 y_folds = np.array_split(y_str3, 5)
323 scores = list()
324 for k in range(5):
325     X_train = list(X_folds)
326     X_test = X_train.pop(k)
327     X_train = np.concatenate(X_train)
328     y_train = list(y_folds)
329     y_test = y_train.pop(k)
330     y_train = np.concatenate(y_train)
331     scores.append(wclf_3.fit(X_train,y_train).score(X_test,y_test))
332     predictions = wclf_3.predict(X_test)
333     # print accuracy
334     print(round(np.mean(predictions == y_test) * 100, 2))
335     # print precision, recall, F1
336     print(metrics.classification_report(y_test, predictions, digits=3))
337     # print confusion matrix
338     print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
    Predicted'], margins=True))
339 # print list of k accuracy scores and mean of scores
340 scores_mean = sum(scores)/float(len(scores))
341 print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')
342
343 # TESTING 3 WAY STANCE (y_str3) WITH ALL FEATURES (X_posi)
344 # divide data into 5 folds to fit and predict
345 X_folds = np.array_split(X_posi, 5)
346 y_folds = np.array_split(y_str3, 5)
347 scores = list()
348 for k in range(5):
349     X_train = list(X_folds)
350     X_test = X_train.pop(k)
```

```
351     X_train = np.concatenate(X_train)
352     y_train = list(y_folds)
353     y_test = y_train.pop(k)
354     y_train = np.concatenate(y_train)
355     scores.append(wclf_3.fit(X_train,y_train).score(X_test,y_test))
356     predictions = wclf_3.predict(X_test)
357     # print accuracy
358     print(round(np.mean(predictions == y_test) * 100, 2))
359     # print precision, recall, F1
360     print(metrics.classification_report(y_test, predictions, digits=3))
361     # print confusion matrix
362     print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
        Predicted'], margins=True))
363 # print list of k accuracy scores and mean of scores
364 scores_mean = sum(scores)/float(len(scores))
365 print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')
366
367 # TESTING 3 WAY STANCE (y_str3) WITH ALL FEATURES (X_no_posi)
368 # divide data into 5 folds to fit and predict
369 X_folds = np.array_split(X_no_posi, 5)
370 y_folds = np.array_split(y_str3, 5)
371 scores = list()
372 for k in range(5):
373     X_train = list(X_folds)
374     X_test = X_train.pop(k)
375     X_train = np.concatenate(X_train)
376     y_train = list(y_folds)
377     y_test = y_train.pop(k)
378     y_train = np.concatenate(y_train)
379     scores.append(wclf_3.fit(X_train,y_train).score(X_test,y_test))
380     predictions = wclf_3.predict(X_test)
381     # print accuracy
382     print(round(np.mean(predictions == y_test) * 100, 2))
```

```
383     # print precision, recall, F1
384     print(metrics.classification_report(y_test, predictions, digits=3))
385     # print confusion matrix
386     print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
        Predicted'], margins=True))
387 # print list of k accuracy scores and mean of scores
388 scores_mean = sum(scores)/float(len(scores))
389 print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')
390 #
391 # TESTING 3 WAY STANCE (y_str3) WITH ALL FEATURES (X_no_acoustics)
392 # divide data into 5 folds to fit and predict
393 X_folds = np.array_split(X_no_acoustics, 5)
394 y_folds = np.array_split(y_str3, 5)
395 scores = list()
396 for k in range(5):
397     X_train = list(X_folds)
398     X_test = X_train.pop(k)
399     X_train = np.concatenate(X_train)
400     y_train = list(y_folds)
401     y_test = y_train.pop(k)
402     y_train = np.concatenate(y_train)
403     scores.append(wclf_3.fit(X_train,y_train).score(X_test,y_test))
404     predictions = wclf_3.predict(X_test)
405     # print accuracy
406     print(round(np.mean(predictions == y_test) * 100, 2))
407     # print precision, recall, F1
408     print(metrics.classification_report(y_test, predictions, digits=3))
409     # print confusion matrix
410     print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
        Predicted'], margins=True))
411 # print list of k accuracy scores and mean of scores
412 scores_mean = sum(scores)/float(len(scores))
413 print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')
```

```
414
415 # TESTING 3 WAY STANCE (y_str3) WITH ALL FEATURES (X_no_med_fin)
416 # divide data into 5 folds to fit and predict
417 X_folds = np.array_split(X_no_med_fin, 5)
418 y_folds = np.array_split(y_str3, 5)
419 scores = list()
420 for k in range(5):
421     X_train = list(X_folds)
422     X_test = X_train.pop(k)
423     X_train = np.concatenate(X_train)
424     y_train = list(y_folds)
425     y_test = y_train.pop(k)
426     y_train = np.concatenate(y_train)
427     scores.append(wclf_3.fit(X_train,y_train).score(X_test,y_test))
428     predictions = wclf_3.predict(X_test)
429     # print accuracy
430     print(round(np.mean(predictions == y_test) * 100, 2))
431     # print precision, recall, F1
432     print(metrics.classification_report(y_test, predictions, digits=3))
433     # print confusion matrix
434     print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
    Predicted'], margins=True))
435 # print list of k accuracy scores and mean of scores
436 scores_mean = sum(scores)/float(len(scores))
437 print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')
438
439 # TESTING 3 WAY STANCE (y_str3) WITH ALL FEATURES (X_acous)
440 # divide data into 5 folds to fit and predict
441 X_folds = np.array_split(X_acous, 5)
442 y_folds = np.array_split(y_str3, 5)
443 scores = list()
444 for k in range(5):
445     X_train = list(X_folds)
```

```
446     X_test = X_train.pop(k)
447     X_train = np.concatenate(X_train)
448     y_train = list(y_folds)
449     y_test = y_train.pop(k)
450     y_train = np.concatenate(y_train)
451     scores.append(wclf_3.fit(X_train,y_train).score(X_test,y_test))
452     predictions = wclf_3.predict(X_test)
453     # print accuracy
454     print(round(np.mean(predictions == y_test) * 100, 2))
455     # print precision, recall, F1
456     print(metrics.classification_report(y_test, predictions, digits=3))
457     # print confusion matrix
458     print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
        Predicted'], margins=True))
459 # print list of k accuracy scores and mean of scores
460 scores_mean = sum(scores)/float(len(scores))
461 print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')
462
463 # TESTING 3 WAY STANCE (y_str3) WITH ALL FEATURES (X_no_word)
464 # divide data into 5 folds to fit and predict
465 X_folds = np.array_split(X_no_word, 5)
466 y_folds = np.array_split(y_str3, 5)
467 scores = list()
468 for k in range(5):
469     X_train = list(X_folds)
470     X_test = X_train.pop(k)
471     X_train = np.concatenate(X_train)
472     y_train = list(y_folds)
473     y_test = y_train.pop(k)
474     y_train = np.concatenate(y_train)
475     scores.append(wclf_3.fit(X_train,y_train).score(X_test,y_test))
476     predictions = wclf_3.predict(X_test)
477     # print accuracy
```

```
478     print(round(np.mean(predictions == y_test) * 100, 2))
479     # print precision, recall, F1
480     print(metrics.classification_report(y_test, predictions, digits=3))
481     # print confusion matrix
482     print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
        Predicted'], margins=True))
483 # print list of k accuracy scores and mean of scores
484 scores_mean = sum(scores)/float(len(scores))
485 print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')
486
487 # TESTING 3 WAY STANCE (y_str3) WITH ALL FEATURES (X_no_task)
488 # divide data into 5 folds to fit and predict
489 X_folds = np.array_split(X_no_task, 5)
490 y_folds = np.array_split(y_str3, 5)
491 scores = list()
492 for k in range(5):
493     X_train = list(X_folds)
494     X_test = X_train.pop(k)
495     X_train = np.concatenate(X_train)
496     y_train = list(y_folds)
497     y_test = y_train.pop(k)
498     y_train = np.concatenate(y_train)
499     scores.append(wclf_3.fit(X_train,y_train).score(X_test,y_test))
500     predictions = wclf_3.predict(X_test)
501     # print accuracy
502     print(round(np.mean(predictions == y_test) * 100, 2))
503     # print precision, recall, F1
504     print(metrics.classification_report(y_test, predictions, digits=3))
505     # print confusion matrix
506     print(pd.crosstab(y_test, predictions, rownames=['True'], colnames=['
        Predicted'], margins=True))
507 # print list of k accuracy scores and mean of scores
508 scores_mean = sum(scores)/float(len(scores))
```

```
509 print('All scores: ', scores, '\n', 'Mean: ', scores_mean, '\n')
```

Appendix E

LIST OF COMMUNICATIONS

E.1 Refereed conferences

2016 | "Weight-sensitive stress and acoustic correlates of disyllabic words in Marathi" *171st Meeting of the Acoustical Society of America*, Salt Lake City, Utah

2016 | "Effects of stance strength and word group on the acoustic properties of the vowel in *um* and *uh* in spontaneous speech in Pacific Northwest American English" *42nd Annual Meeting of the Berkeley Linguistics Society*, Berkeley, CA

2016 | "Discursive patterns of *um* and *uh* in spontaneous speech in Pacific Northwest American English" *90th Annual Meeting of the Linguistic Society of America*, Washington D.C.

2015 | "Weight-sensitive stress in disyllabic words in Marathi and the case of the high vowels" *31st Northwest Linguistics Conference*, Victoria, BC

2014 | "Categorization issues and POS tagging of the marker 'so' in oral English" *12th ESSE conference*, Kosice, Slovakia

2012 | "Form and Function of Connectives in Oral Narratives" *16ème colloque d'anglais oral de Villetaneuse « L'anglais oral mondialisé ? »*, Villetaneuse, France

E.2 Presentations

2016 | "Tools for teaching French pronunciation : teaching strategies and awareness of non-standard language. Capstone Project for the University of Washington Graduate Certificate in Second and Foreign Language Teaching." *French and Italian Studies Graduate Student Colloquium*, University of Washington, Seattle, WA

2016 | "Discourse and acoustic patterns in *um* and *uh* in spontaneous speech and the effect of Stance" *Linguistic Graduate Student Colloquium*, University of Washington, Seattle, WA

2015 | "Investigating and Testing Weight-sensitive Stress in Marathi" *Guest lecture, Linguistics Phonology II*, University of Washington, Seattle, WA

2014 | "Linear Mixed Effect Models" *University of Washington Phonetics Lab Meeting*, Seattle, WA

2013 | "Pitch Resynthesis" *University of Washington Phonetics Lab Meeting*, Seattle, WA

2013 | "Corpus interoperability and spoken diachronic databases : the NECTE-DECTE corpora" *Université Paris Diderot*, Paris, France

2013 | "Dialects of French and English, emphasis on Breton and Indian English" *Guest lecture, Introduction to Phonetics*, University of Washington, Seattle, WA

2012 | "What is Phon ?" *University of Washington Phonetics Lab Meeting*, Seattle, WA

2012 | "Interface between POS tagging, semantic categorization and the prosodic features of the marker 'so' : a multilevel study" *CLILLAC-ARP laboratory Poster Session*, Paris, France

E.3 Article

Le Grézause, E. (2015). Investigating Weight-Sensitive Stress in Disyllabic Words in Marathi and its Acoustic Correlates. *University of Washington Working Papers in Linguistics*, 33, 33–52.

E.4 Scholarships and Certificates

2016 | Excellence in Linguistics Research Graduate Fellowship
University of Washington, Seattle

2016 | Graduate Certificate in Second/Foreign Language Teaching - French
University of Washington, Seattle

2015 | University of Washington Travel Grant

2014 | Software Carpentry Instructor for Python, R, The Unix Shell