Dialogical Signals of Stance Taking in Spontaneous Conversation

Leanne Rolston

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

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Gina-Anne Levow, Chair

Richard Wright

Emily M. Bender

Program Authorized to Offer Degree:
Department of Linguistics

University of Washington

Abstract

Dialogical Signals of Stance Taking in Spontaneous Conversation

Leanne Rolston

Chair of the Supervisory Committee:
Gina-Anne Levow
Department of Linguistics

This is one of the first computational studies to investigate dialogical aspects of stance taking in spontaneous, spoken dialogue with a focus on lexical similarities. In any dialogic interaction, each speaker influences the others' lexical choices and aspects of their grammatical style (Brennan, 1996; Niederhoffer and Pennebaker, 2002).

For this study, I leverage two distinct corpora. The ATAROS corpus (Freeman et al., 2014; Freeman, 2015) contains a series of task-oriented, collaborative tasks recorded in a controlled laboratory environment. The other corpus is drawn from a United States Homeland Security Subcommittee hearing regarding the 2007 - 2008 financial crisis. As such, it represents stance taking in an inherently adversarial environment, where high-stakes, real-world issues are being discussed. Both are annotated at the spurt-level with a 3-way stance strength annotation.

I will show, through various experimental studies, that speakers show different patterns of dialogical behaviour when expressing stance versus when they are not; they show a higher level of engagement, as demonstrated by the use of similar or related terminology, and the rate of convergence in linguistic style, as measured by function word use, is also higher when expressing stance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to all the office mates who let me bounce ideas off them when I was probably not making sense to anyone but myself (Thanks Ducks! Kristen Howell primary among you!), to the folks online who tolerated my rants and attempts to do the same (Brent Woo & Molly FitzMorris), and my gmail drafts folder that now holds many potentially embarrassing e-mails I was in the midst of writing for help whose writing actually helped me figure out the problem.

Also, thank you to the coffee shops who offered me hours of free WiFi.

In the UW Linguistics Department, I'd like to thank the front office staff, past and present, for all the talk and candy; my advisor, Gina for her un-ending patience with me; my fellow students for the company; and all faculty and staff for letting me share this part of my life with you.

Outside of UW, I'd like to thank the schools and camps that kept my kids occupied and safe so I could focus on my schooling: Bright Horizons - Kirkland for those early years, Wilder Elementary and Les Lilas for the later; and particularly Studio East and Eton School for their online programs so that I could finally finish.

Thanks to my family for the love and support.

And, finally, thank you to Mike Mueller of Row House Redmond for the loan of the ERG during the COVID-19 shutdown. You have no idea how often I rowed myself out of writer's block.

DEDICATION

Most of this was written during the COVID-19 shelter in home order while home schooling my two sons. Thank you to Hugh and Ian for whatever time you could give me, and to my husband Adam for the years of support.

Chapter 1
INTRODUCTION

Stance, and more specifically *stance taking*, is the expression of one's personal thoughts and attitudes toward something (Haddington, 2004). Stance can be overtly stated, or inferred from the surrounding dialogue. For the interpretation of stance, three aspects must be identified: the stance taker, which is frequently the speaker, the stance object, that is the entity or proposition to which the expression of stance refers, and the party to which the expression of stance is expressed, the fellow interlocutors. Because of this, stance is an inherently dialogic and social act (Du Bois, 2007); any expression of stance can be understood to be either directly or indirectly in response to something mentioned in the discourse history or in response to some socio-cultural norm (Du Bois, 2007). The term *subjective* is sometimes used to describe the nature of these attitudes; this contrasts with the term *objective* or factual, and describes thoughts and feelings that are not "open to objective observation or verification" (Quirk et al., 1985).

Stance can be expressed in a variety of ways, the most obvious way being the direct expression of how one feels. Stance can also be expressed through the use of *value-laden word choice* (Biber et al., 1999) or *expressive subjective elements* (Wilson et al., 2004). Contrast the terms *that man* and *that jerk*; the former is a neutral term while the latter reveals the opinion of the speaker toward the person in question. Sarcasm is another means to express stance (Riloff et al., 2013), however familiarity with the speaker and their circumstances are required for an accurate interpretation (Riloff et al., 2013). This is not unique to sarcasm, however; many linguistic elements are subjective only in specific contexts. Subjectivity is not a property of words, but word senses (Wiebe and Mihalcea, 2006). Contrast the word *quack* when it refers to the sound a duck makes, or when it refers to a medical professional.

What is meant by *dialogical*, then? A dialogue is more than just topically relevant speaking turns spoken by multiple speakers. It is a highly coordinated activity. Participants must share **common ground**, that is, a set of mutual beliefs, assumptions, and knowledge about the content of the dialogue, the environment in which it is happening, and the participants, both those who are actively speaking and those who may be just observing. This common ground develops and changes with each contribution to the dialogue through a process known as **accumulation** (Clark and Schaefer, 1989). As speakers contribute information to the dialogue, speaking partners are given the opportunity to accept it into the common ground or reject it. Acceptance is signaled through overt acknowledgement (continuers or back-channels), speakers moving on to the relevant next turn, or simply by showing continued attention (Clark and Brennan, 1991). Where the information is not accepted, whether it was not heard or understood by the audience, or understood and the other speakers do not agree, a repair is initiated. This process, where speakers establish the common ground, and seek clarification when they do not, is known as **grounding** (Clark and Brennan, 1991).

1.1 Statement of the Problem

Much of the work on stance and subjectivity focuses on the words or grammatical structures used to express it. Unigram or other n-gram features were shown to be successful for the classification of the semantic orientation (positivity or negativity) of online reviews (Pang et al., 2002) and classifying whether a post is "for" or "against" a specific topic in online debates (Somasundaran and Wiebe, 2010; Anand et al., 2011); these, were, however, very domain and topic specific (Liu, 2010).

Annotation projects revealed the wide variety of grammatical forms and structures used to express subjectivity. Documents from the Wall Street Journal portion of the Penn Treebank (Marcus et al., 1993) were annotated with a sentence-level judgement of *subjective* or *objective* (Bruce and Wiebe, 1999; Wiebe et al., 1999). In a followup study, Wiebe (2000) had the same annotators specify the specific subjective elements within these sentences. Word level analysis of inter-annotator agreement showed only moderate agreement (Cohen's kappa coefficient of

0.42) (Wiebe et al., 2004) as opposed to the sentence-level agreement of 0.76 (Wiebe et al., 1999).

For the MPQA Corpus, Wilson and Wiebe (2003) used English language articles from the U.S. Foreign Broadcast Information Service (FBIS), and had annotators mark the span of text that expressed subjectivity. Once the span was identified, they were tasked with assigning other attributes to the span, such as a strength measurement. They again found that though annotators identified the same spans, they differed in the exact boundaries they marked. A pairwise precision measurement for word-level inter-annotator agreement across the entire data set averaged 72%.

Both annotated corpora were used extensively in experiments to learn methods of finding subjectivity clues in unannotated data. Adjectives were identified as being correlated with subjectivity (Bruce and Wiebe, 1999) and were used as a seed set to identify higher quality adjective features using a distributional similarity measure; maximum precision, however, was found only after pruning the set of candidates and adding the lexical semantic features of gradability and polarity (Wiebe, 2000; Wiebe et al., 2004). Additional non-word based features such as the fact that a term was low frequency and the density of subjectivity clues were also found to be reliable clues of subjectivity (Wiebe et al., 2004).

A close analysis of the annotations in the MPQA Corpus shows the breadth of linguistic structures that were used to express subjectivity. While the prototypical means to express subjectivity directly is through the use of a verb, such as *think* or *like*, the data showed that verbs were used only 54% of the time; the remainder were expressed using nouns (32%), adjectives (8%), and adverbs (6%) (Wiebe et al., 2005). Additionally, many of the terms that were annotated as holding a subjective meaning were also used in an objective statement elsewhere in the corpus (Wiebe et al., 2005). Since subjectivity is a property of word *senses*, not of words themselves (Wiebe and Mihalcea, 2006), the learned patterns and lexical clues are only **potential subjective elements**; only in context can it be determined whether they signal subjectivity or not.

There are also acoustic clues to subjectivity. Somasundaran et al. (2006) annotated subjectivity in multi-party meetings; when annotators were given the audio recordings along with the transcripts, inter-annotator agreement increased on average 0.18 Kappa points over annotation based on transcripts alone. Freeman (2015) developed the ATAROS corpus to investigate the acoustic correlates of stance taking, and found that stance strength correlates with intensity and pitch. A close investigation into the token *yeah* showed that it was used to express positivity only 68% of the time; the remaining uses were neutral or negative (Freeman et al., 2015).

The acoustics, however, do not tell the whole story, either. Raaijmakers et al. (2008) compared prosodic features to word, character, and phone n-grams, and while the combination of all four features yielded the best accuracy, recall, and precision, as individual features, prosody was successful in distinguishing positive from negative subjectivity only. Levow et al. (2014), using the ATAROS corpus, compared the acoustic correlates found by Freeman (2015, 2019), with speaking style features, such as spurt duration and rate of disfluencies, and lexical features, such as unigrams, and found that unigrams, when punctuation was included, were the most discriminative features for binary stance detection and a four-way stance strength detection. These unigrams are given in Table 1.1.

yeah	okay	um	hm
need	maybe	important	good
the	this	could	but
?	,	!	

Table 1.1: Discriminative Unigrams given in Levow et al. (2014)

As this list shows, at least in a dialogical setting, clues to stance taking are not necessarily terms found in a subjectivity clues lexicon. The subjectivity of an utterance seems to be expressed in how it is said as much as in the content of the utterance.

This study seeks to answer whether speakers demonstrate different patterns of dialogical behaviour when expressing stance versus when they are not. It uses the ATAROS corpus (Freeman, 2015). Starting with speakers' word choices, I first demonstrate how stance strength is related to the strength of the words used to express it along the emotional dimensions of *valence*, *arousal*, and *dominance*; I also investigate whether speakers are influenced by the strength of the words used by their speaking partner, specifically whether one speaker's use of language strong in any of these dimensions influences the strength of the words the other speaking partner uses along the same dimension. Next, I show that speakers align with each other in their terminology in stance-laden speaking turns more than those that do not express stance using a speaking turn-level document similarity measurement, Word Movers Distance (Kusner et al., 2015), which I use to quantify dialogic resonance. Then I give examples where these resonances demonstrate a shared conceptualization of the topic under discussion. Finally, I show that speakers also converge on a less conscious level through an analysis of function word use when they are expressing strong stance moreso than weak or no stance using the Linguistic Style Coordination measurement introduced by Danescu-Niculescu-Mizil et al. (2012).

1.2 Paper Layout & Contributions

The remainder of this paper is organized as follows. Chapter 2 provides some background on the computational fields of stance and subjectivity detection. Chapter 3 describes the ATAROS corpus and its development. Experimental work starts in Chapter 4 with the investigation into the correlation between word strength and stance strength. Next, Chapter 5 looks into terminological similarities as a measurement of engagement. Finally, Chapter 6 investigates the relationship between Linguistic Style Coordination (Danescu-Niculescu-Mizil et al., 2012) and stance strength.

The interplay between stance and dialogue is complicated. While there has been a lot of work trying to itemize the means used to express stance at the word or phrase level in single threaded texts, relatively little computational work has been done on stance taking in

dialogue. The results of Levow et al. (2014) show that it is more complicated than just the use of stance-laden lexica or grammatical structures. This study will investigate patterns of stance taking in dialogical interaction, and how these patterns differ when speakers are expressing stance versus when they are not.

Chapter 2
BACKGROUND

Most of the work on stance and subjectivity detection involves the classification of single-threaded texts such as news articles or reviews. Initially, texts were classified at the document level. Later methods targeted classification at the sentence or phrase level. These have traditionally revolved around finding lexical clues or grammatical structures that can be used to identify subjective statements, and optionally whether these statements express positive or negative sentiment toward the target, which is known as polarity or semantic orientation.

Studies relating to stance taking in a dialogical setting focused on multi-party meetings, online debates, or United States Congressional floor debates. While these were able to incorporate features relating to the multi-party nature of the interactions, the majority of the features continued to focus on the lexical content in isolation.

The results of Levow et al. (2014) described in Chapter 1 make it clear that, in addition to the grammatical and lexical clues available from the content of an utterance, there must also be signals of stance taking in the way speakers interact. The influence speakers have on their speaking partner has been studied in computational research on entrainment.

2.1 Subjectivity Lexicons

The first major corpus investigations into the marking of stance in English were undertaken by Biber and Finegan (1988, 1989) using spoken and written texts of various genres. Their focus was initially on stance adverbials, before expanding to other grammatical structures. Using a list of stance-marking lexical items based on work done by Quirk et al. (1985), they used document clustering and subsequent analysis to determine that the relationship between stance categories and the grammatical means of marking stance was not as straightforward as it was initially thought to be.

Early computational work focused on the classification of texts as subjective or objective (Yu and Hatzivassiloglou, 2003). This was seen as necessary step for tasks such as information extraction and document classification (Wiebe et al., 2004). Other work focused on the classification of inherently subjective texts such as reviews and determining whether they express positive or negative sentiment toward a target (Turney, 2002; Hu and Liu, 2004). Much of this work focused on word-based features such as unigrams (Pang et al., 2002; Somasundaran et al., 2007). These were, however, very domain and topic specific (Liu, 2010).

While annotation projects, such as those mentioned in Chapter 1 revealed a broad spectrum of subjectivity clues in English, work focusing on a specific grammatical category also continued. Hu and Liu (2004), focused on the use of adjectives in product reviews, using the co-occurrence of an adjective with a product name or feature as a subjectivity clue; the semantic orientation for unknown adjectives was found through leveraging semantic relationships (synonymy, antonymy, hyponymy, and derivation) in WordNet (Fellbaum, 1998). The results of this and subsequent works were released as the Opinion Lexicon. Jindal and Liu (2006) focused on the use of comparatives, such as *-er* adjectives, verbs like *outperform* and terms such as *number one*. Their experimentation found that the co-occurrence of these comparatives alongside words such as *whereas*, *however*, *although* yielded the highest precision among all experimental conditions.

Applying methods from Information Extraction, Riloff and Wiebe (2003) focused on lexico-syntactic extraction patterns, for example *noun preposition <np>* to capture the phrase *opinion on <np>*. This work was furthered by Wiebe et al. (2004), who, in addition to applying these lexico-syntactic patterns toward expanding their inventory of subjective clues, also found that terms used with low frequency were often subjective, and that subjective texts used more low frequency terms overall. This, and other work was combined into the Subjectivity Lexicon (Wilson et al., 2005b) which is used in the Opinion Finder (Wilson et al., 2005a).

In developing the Arguing Lexicon, Somasundaran et al. (2007) identified instances of arguing in the AMI Meeting Corpus (McCowan et al., 2005) based on dialogue acts, then extracted the n-grams used to express these acts. This was intended to expand and create a more general arguing lexicon. From this, they discovered that closed class words such as modal verbs, adverbs, and conjunctions were reliable signals of argumentation in English.

Some of the lexicons mentioned thus far were released, however it should be noted that their contents are only potentially subjective elements. Wiebe and Mihalcea (2006) show that subjectivity is not the property of a word, but a word sense. As an example, the word *quack*, when referring to a doctor is an expressive subjective element. When it refers to the sound a duck makes, it is not.

While most of the work mentioned so far focuses on finding direct mentions of opinion, Choi and Wiebe (2014) do opinion inference through the use of words that reveal evaluation on the effect of the action described. These are words like *stimulate* expressing positive evaluation toward the entity that is being stimulated, and *curb* revealing negative evaluation toward the entity that is being curbed. This lexicon was released as the +/-EffectWordNet.

2.2 Sentiment Lexicons

In addition to lexical subjectivity clues, many of the approaches also use word-level sentiment strength lexicons. These initially came from the field of psychology. The Affective Norms for English Words (ANEW) (Bradley and Lang, 1999) used a seven-point scale of icons to have subjects rank English words along the dimensions of valence (negative $\leftrightarrow$ positive) , arousal (bored $\leftrightarrow$ excited), and dominance (submissive $\leftrightarrow$ dominant);[1] this corpus complements the International Affective Picture System (IAPS) and International Affective Digitized Sounds (IADS).

Language Inquiry and Word Count (LIWC) (Pennebaker et al., 2007, 2015; Tausczik and Pennebaker, 2010) is a text analysis program stemming from studies dating back to the

[1]These terms will be further described in Chapter 4

1980's. The first iteration tied the language used by hospital patients to their recovery; later LIWC was used to map word use to general psychological state, highlighting social coordination, dominance, honesty, and deception. The initial studies had manual evaluation of the text by trained annotators. This was gradually automated with human-curated dictionaries, and the classification of words into 80 categories. These dictionaries were iteratively built and updated between 1992 and 1994, the streamlined in 1997 and 2007, and 2015. The later iterations were tailored for text from the Internet, and adopted many forms of Internet shorthand.

SentiWordNet 3.0 (Baccianella et al., 2010) automated the process of assigning scores to words using semantic relationships in WordNet (Fellbaum, 1998). Starting with seven "paradigmatically positive" and seven "paradigmatically negative" seed terms, new candidate terms were added based on relationships between these seed terms and other synsets using relationships that preserve polarity (for example "also-see") or invert polarity (for example direct antonymity) within a specified radius.

Many other sentiment lexicons were scored using crowd sourcing. The Norms Lexicon (Warriner et al., 2013) selected words based on frequency calculated from the SUBLEXT-US corpus (Brysbaert and New, 2009) with calibrator and control words chosen from the ANEW corpus (Bradley and Lang, 1999). Annotators were asked to annotate words on one of three emotional dimensions (valence, arousal, or dominance) using a 1 – 9 scale, where 1 represents the strong end of the scale (happy/excited/controlled) and 9 represents the low end (unhappy/calm/in control); they were instructed to rate a word a 5 if the word does not evoke that particular emotion. Valence and arousal scores were reversed to represent a low-to-high scale, which is more intuitive. The resulting per-word score represents the mean of the participants' rating for each dimension.

The Evaluative Lexicon (Rocklage and Fazio, 2015) focused on adjectives. Ninety-four adjectives were selected because they were deemed to be evaluative in nature; had an obvious, unambiguous, positive or negative denotation; represented a wide range of valences and emotions; and were applicable across multiple domains. Crowdsourced annotators were

asked to judge a word's valence and emotionality on a 0 – 9 scale. Each word's valence score was averaged across users, while the extremity score was the deviation from the midpoint of the valence score, averaged across users. This results in a range of scores from 0 – 5 for extremity and 0 – 9 for valence and emotionality.

The National Research Council of Canada has created several scored sentiment lexicons. The NRC Affect Intensity Lexicon (Mohammad, 2017) scores terms along the dimensions of anger, fear, sadness and joy, while the NRC-VAD Lexicon (Mohammad, 2018) scores words along the dimensions of valence, arousal, and dominance. Both works build upon previous word lists, starting with the NRC Emotion Lexicon (Mohammad and Turney, 2013), which is a non-scored lexicon that simply lists words and their associated dimension. These terms were supplemented with words that have a high rate of co-occcurrence with emotion-word hashtags as determined by a pointwise mutual information (PMI) measure calculated for each emotion from the Hashtag Emotion Corpus (Mohammad, 2012). Annotators were presented with word 4-tuples and were asked to select the word that most and least evokes the dimension in question. Scores were calculated using Best Worst Scaling (BWS) (Louviere and Woodworth, 1991) which has been shown to yield more reproducible and therefore more consistent rankings (Mohammad, 2017). The final score was the proportion of times a word was chosen as most evoking the dimension minus the proportion of times a word was chosen as least, transformed into a 0 – 1 scale using a linear function. The NRC-VAD Lexicon is used for the the word-level strength scores used in Chapter 4.

While many of these sentiment lexicons have been integrated into studies of stance-taking and subjectivity detection, very few studies have directly addressed the relationship between stance strength and lexical strength. Additionally, most studies use a single score, most analogous to the dimension of valence. This study marks one of the first to directly investigate lexical strength along the dimensions of valence, arousal, and dominance, and stance strength.

2.3 Stance Taking in a Dialogical Setting

Research focusing on stance taking in a dialogical setting uses online debates, political discussion, and multi-party meetings.

Somasundaran and Wiebe (2010) investigated whether sentiment and arguing based features were useful for the classification of "side" (for or against) in ideological online debates. Arguing features were based on n-gram matches with items from a subjectivity lexicon. Similarly, a sentiment lexicon was queried for the semantic orientation (+, -, or neutral) of all the words in a sentence; the dominant orientation, accounting for negating expressions, was assigned to the sentence as a whole. They found that a combination of arguing and sentiment features outperformed both types of feature subclasses, and unigram-based features on their own.

Anand et al. (2011) investigated the use of dialogical features in online debates. They found that features such as the use of quotations, second person pronouns and negation were useful for identifying dialogical features, such as rebuttals, however when they are applied to the classification task of whether the post is for or against the topic, they did not improve upon a classifier using post-internal features, such as n-grams and sentence length.

Thomas et al. (2006) studied stance taking in United States Congressional floor debates. They classified speaking turns as expressing opposition or support for a specific topic using unigram features as well as the speaker's references to other speakers in the discourse; whether these speaker references marked agreement or disagreement was determined by whether they voted the same way in the floor vote. Interestingly, they found the best classification results when the data from multiple speaking turns was concatenated into a single unit, indicating that dialogical markers of stance taking may aggregate across speaking turns.

Somasundaran et al. (2006) annotated the ISL meeting corpus (Burger et al., 2002) with a speaking turn level sentiment category (positive/negative sentiment, positive/negative arguing), and a four-point intensity level using meetings that focused on task-oriented discussion. An analysis of the annotations showed that using only the transcriptions, only positive arguing and sentiment were reliably annotated. The audio was required for annotators to detect

negativity.

Wilson (2008) adapted the MPQA Annotation Scheme (Wilson and Wiebe, 2003) for a multi-party meeting setting and used it to annotate the AMI Meeting Corpus (Carletta et al., 2005) to form the AMIDA Corpus. When comparing the subjectivity annotations and the original dialogue act tags, which focus on speaker intent, it was discovered that there is not a perfect correspondence between the two; many utterances marked as *assessments* were not considered subjective, while many statements marked with the dialogue act *inform* were. These annotations were used by Raaijmakers et al. (2008) who found that word and character n-grams were the best individual features for classifying speaking turns as subjective versus objective, and for a semantic orientation (positive or negative).

While these studies have a dialogical or multi-party aspect to them, the register is overall very formal. Internet forums and Congressional floor debates, may be better characterized as interactive monologues. Speakers are given the "floor" and are able to articulate their thoughts largely without interruption. Even in meetings that are more interactive in nature, the environment, and the size of the group allows for more traditional turn taking behaviour. The ATAROS corpus, in contrast, features two people working together on a task for which they were given very little opportunity to prepare. The smaller group size, and the fact that speakers are allowed to define aspects of the task and task items themselves allows for a more interactive dialogue; speaking turns are overall shorter, many contain incomplete thoughts, and overlapping and interrupting speech is common.

2.4 Subjectivity & Sentiment Strength

Several studies use a scored lexicon to represent sentiment strength, if only to help make a 2-way distinction between positive and negative evaluation. Turney (2002) used a measure of Pointwise Mutual Information and Information Retrieval (PMI-IR) with the terms "excellent" and "poor" to quantify the semantic orientation of phrases containing adjectives and adverbs and adjacent contextual words. This score was reported to be positively correlated with the "star-based" rating used by most review sites, however the strength of the score

was not otherwise addressed. Other tools, such as the Semantic Orientation CALculator (SO-CAL) (Taboada et al., 2011) and SentiStrength (Thelwall, 2013) use a lexicon-based approach in which a sentiment lexicon is built, and each word is assigned a numeric value representing its prior polarity, which is the semantic orientation of the word independent of context (Taboada et al., 2011). The semantic orientation of the text is calculated from the semantic orientation of the words making up the text, factoring in grammatical aspects such as negation, intensifiers, and hedges.

In terms of the strength of a subjective statement, the Multi-Perspctive Question Answering (MPQA) corpus (Wilson and Wiebe, 2003) has a strength value representing the intensity with which the opinion is expressed, along with the source and target of the opinion. There were four intensity values: *neutral*, *low*, *medium*, and *high*; a *neutral* value refers to a statement that does not express any opinion. Classification experiments on this corpus by Wilson et al. (2004) showed classification of subjective strength to be successful using a combination of lexical, grammatical, and dependency features.

The ATAROS corpus (Freeman et al., 2014; Freeman, 2015) was developed to measure the acoustic correlates of stance taking in spontaneous speech. Using manual stance strength annotations which will be explained further in Chapter 3, it was found that intensity and pitch correlates with stance strength, that is, stronger stances are spoken more loudly and with greater pitch changes than less stance-laden utterances.

The studies above address either sentiment strength, or subjective strength. This study marks one of the first to address both.

2.5 Entrainment

In any dialogic interaction, speakers begin to converge in aspects of their speech, whether it be acoustic-prosodic, lexical, grammatical, or syntactic. The influence speakers have on each other is broadly categorized as entrainment. Entrainment has also been reported to correspond with measures of speaking compatibility and task success, though many studies have also found evidence of entrainment in dialogues where the participants self-reported a

lack of rapport with the other speakers, or difficulty in completing the task.

2.5.1 Lexical

Garrod and Anderson (1987) showed that speakers coordinated on the terminology used to describe movement through a grid pattern. The task had speakers direct each other to move a marker from a starting point to a target location. They concluded that the fact that speakers came to an agreement on the use of terms such as *node* versus *box* or *bottom* versus *last* showed that speakers had a shared conceptualization of the task.

Brennan and Clark (1996) note that, while there are many terms that can be used to refer to an item or a concept, only a limited number are used within a single conversation. In a series of experiments, they had dyads complete a series of matching tasks. For the first set, the items represented on the cards were each of a unique base category, such as a shoe, a dog, and a car. Here, they found speaking partners referred to the items by the the base category. The second set of cards had these same items, plus additional cards of these same base categories. Here, the speakers used a more specific description of the items to distinguish them from others of the same base category. When given the set of cards with unique base categories for the third task, speakers continued to use the more specific term established in the second task, rather than revert to the more general task used for the first. This, they describe as a *conceptual pact* between speakers.

In a follow up study, Brennan (1996) tested whether these conceptual pacts would extend to the interaction between a human and a computer, either spoken or in text, and how the type of correction affects this adoption of terminology. Using a Wizard-of-Oz scenario, participants were asked to do a database query. The system then performed a confirmation step varying whether the same term or a different terms was used. For example if the user used the term *college* in the query, the confirmation would use the term *school*. They found that a speaker is more likely to adopt the system's terminology when the system used an *exposed correction* ("By college, did you mean school?") versus an *embedded correction* ("The school attended was").

Nenkova et al. (2008) studied entrainment of high-frequency words using the Columbia Games Corpus (Gravano and Hirschberg, 2011), a set of competitive, collaborative tasks. Their measurement was a comparison of the proportional use high frequency words between speaking partners. They found that a high degree of entrainment in high frequency words corresponds to dialogical cohesion, and task success.

2.5.2 Syntactic

Branigan et al. (2000) showed that speakers coordinate in grammatical form by using an experiment where participants were asked to describe actions involving ditransitive verbs. This study contrasted the grammatical structures *gives <patient> to <beneficiary>* and *gives <beneficiary> <patient>*, and found that, speakers were likely to match their speaking partner in grammatical structure, even when the verb, patient, and beneficiary roles were filled by different words.

Niederhoffer and Pennebaker (2002) studied entrainment in online chatrooms. In addition to finding entrainment in turn length, speakers also entrained in word type use, specifically in function words, a phenomenon they named **Linguistic Style Matching**. This was found to happen in conversations even where the participants self-reported that they did not "click" with the conversation partner.

Danescu-Niculescu-Mizil et al. (2012) devised a quantification of Linguistic Style Matching which they called **Linguistic Style Coordination**. It is the probability of the use of a function word category given its use in the immediately preceding turn. It was found to be correlated with power differentials between speaker groups using a corpus of Wikipedia meta discussion and US Supreme Court arguments. Chapter 6 will be using this measurement to investigate the relationship between linguistic style and stance strength.

2.6 Summary

Clues that a speaker is expressing something subjective can lie in the content of the utterance, or, it seems in the interaction between speakers. Subjectivity and sentiment lexicons can be

used to identify many instances of stance taking, however, not all expressions of subjectivity use overt clues. In studying how speaker interaction differs when expressing stance versus not, I hope to find some of the more subtle clues to stance taking.

Chapter 3
THE ATAROS CORPUS

Since the majority of the research on subjectivity focuses on single threaded texts, there are few resources available to study subjectivity in multi-party, spoken interactions. Stance-rich genres that are available, such news media interviews or political debates, tend to involve speakers in a highly emotive, combative state, which makes it difficult to establish whether their speaking behaviour is a product of stance taking or emotion. To determine which speaker tendencies are a product of stance taking, and to find the more subtle cues, a more controlled environment is required.

The ATAROS (Automatic Tagging and Recognition of Stance) corpus was developed by Freeman (2015) to study the acoustic correlates of stance taking in unscripted, spontaneous, English speech. It was designed as a series of collaborative, goal-oriented tasks in which participants are instructed to discuss the task, on the way to coming to an agreement. All participants were native speakers of American English from the Pacific Northwest dialect region, between the ages of 18 and 75. Speakers were paired with another speaker of roughly the same age, who they did not know prior to the session. To mitigate any effects of gender imbalance, gender combinations were controlled for. Recording sessions lasted roughly one hour, and took place at the University of Washington in a sound attenuated booth. For this study, data from 30 dyads was used, comprised of 37 women and 23 men.

The following sections will describe the tasks, transcription and stance strength annotation. Additionally, high-level trends in stance taking found in the corpus will be investigated.

3.1 Task Design

After demographic information was collected, each dyad completed a set of collaborative tasks. These tasks were divided into two groups, with each group having its own core vocabulary, which can be found in Appendix A. Each task group starts with a find-the-difference style task that is used to familiarize the participants with the core vocabulary and elicit stance-neutral utterances of the key vocabulary terms. These are followed by the tasks in which we expect to find many instances of stance taking.

The first stance-eliciting task is the inventory task, referred to as **ATAROS 3I** throughout this study, in which the participants negotiate the layout of a department store. It was designed to elicit weak stance and relatively low levels of engagement. Participants were given a felt board representation of a store layout with multiple aisles and a box of cards with task items written on them. The discussion revolved around the placement the cards on the board, and the task was considered completed when both partners were satisfied with the layout.

The other stance-eliciting task is the budget task, referred to as **ATAROS 6B**, in which the participants are asked to balance a municipal budget. It is intended to elicit stronger stance and higher levels of engagement. For this task, participants were given a list of county departments, and under each department was a list of candidate items considered for defunding. The group's task was to discuss each item and agree which ones to keep and which ones to cut from the budget.

Figure 3.1 shows the task duration for each dyad. A horizontal dashed line shows the task mean. Note that although these look close, these are separate values; the 3I task lasted, on average 10m16s, with a standard deviation of 4m58s while the 6B task mean was 10m35s with a standard deviation of 7m19s.

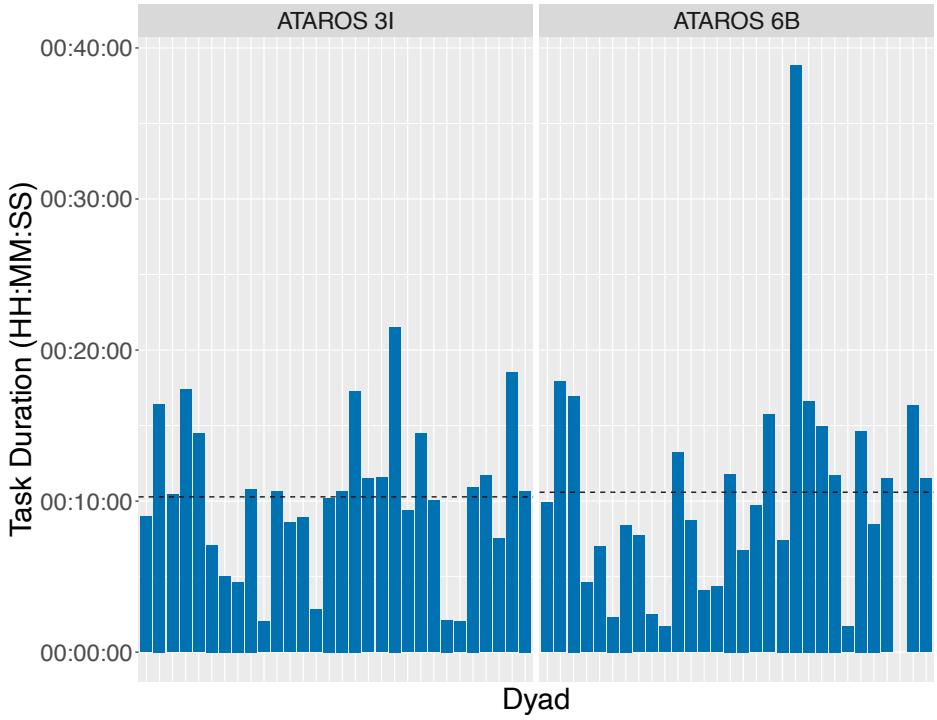


Figure 3.1: Task Duration for the ATAROS Tasks

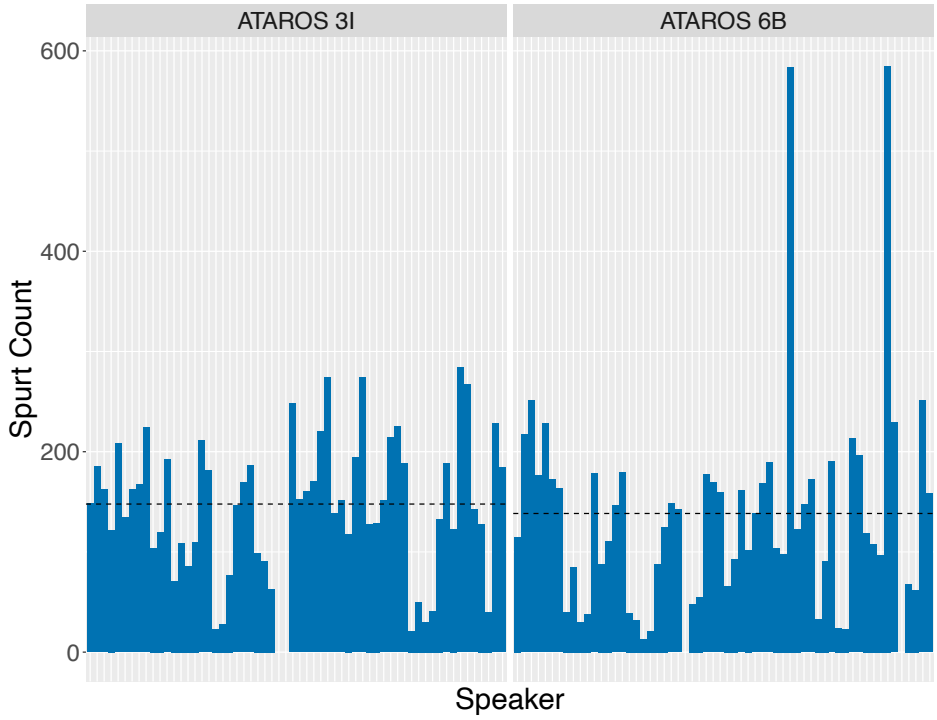
3.2 Transcription

After the recording session, the audio files were divided, one file for each task, for transcription and stance strength annotation. Praat (Boersma et al., 2002) TextGrids were created for each file, and each speaker in the dyad was transcribed on a separate interval tier. The transcription tier was divided into **spurts**, where a spurt is defined as speech surrounded by at least 500ms of silence, following the model used by Shriberg et al. (2001). Each spurt was transcribed orthographically. Annotators were instructed to use common American English spelling with the exception of common shortened versions of words (for example "cuz" for because), phonological contractions (gonna, sorta), discourse markers (uh-oh, hm), and a pre-determined set of common vocalizations (meh, psh). Pauses of less than 500ms were marked with an ellipsis (..), filled pauses marked with "um" to indicate audible nasality or "uh" to indicate the absence of nasality. A dash (-) was used to indicate an abrupt pause,

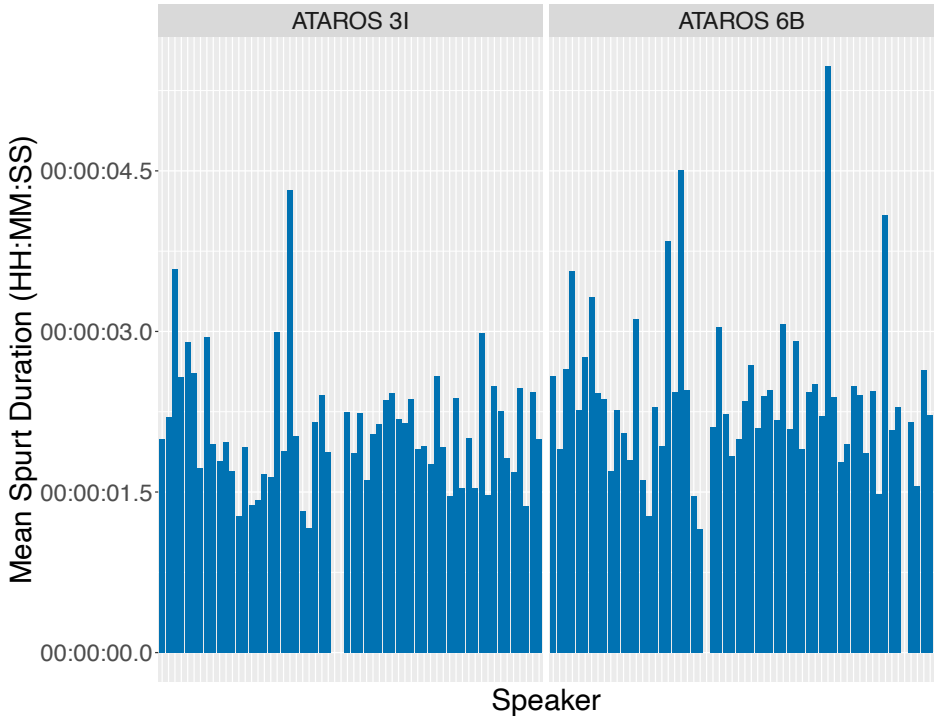
an abrupt continuation of speech, or an utterance broken off mid stream.

Using these orthographic transcriptions, automatic phone alignment was inserted into another tier of the TextGrid, one for each speaker. This is described in detail in Freeman (2015).

Figure 3.2 shows the count of spurts for each participant in the ATAROS 3I and 6B tasks, and their average spurt duration. A horizontal line on Subfigure 3.2a shows the average spurt count for each task; these are close, but not identical. The average spurt count for the 3I task was 147.8, and for the 6B task was 138.4. Subfigure 3.2b shows the per-speaker mean spurt duration, with the cross-speaker average, which is not marked on the figure, at 2.09 seconds for the 3I task, and 2.41 for 6B. Gaps in these figures are due to either too little or incomplete data for a speaker. Three speakers registered fewer than 30 content-rich spurts (i.e. spurts with textual content) within a task, and were therefore too small a bar to appear on the scale and another dyad was represented in one task and not the other.



(a) Per-Speaker Spurt Count



(b) Per-Speaker Average Spurt Duration

Figure 3.2: Per-Speaker Spurt Count and Average Duration

This and Figure 3.1 may seem slightly contradictory in that the 6B task took more time, yet had fewer spurts. This is due to the nature of the tasks. The 3I task was largely a visual task, with the speakers using a felt board to view and organize their layout. As a consequence, many of the suggested article placements were expressed using terms like *here* or *there* which required disambiguation, such as in the following:

A: Rolls of duct tape. This gotta be over here someplace. (2)

B: Yeah, how bout duct tape down there? (1)

A: Right there? (1)

B: Over there with, uh - Okay. That's good. Coulda been with power cords, but that's okay. (1)

Note that speaker B also provided a less ambiguous descriptive location, *with the power cords*. This is similar in nature to the references in the ATAROS 6B corpus, which did not have a visual component.

3.3 Annotation

After transcription and phonetic alignment, annotators were tasked with assigning a stance strength judgement to each spurt; these were inserted into the TextGrid on another interval tier, again one per speaker. This judgement was based on both textual content and prosodic features (Freeman, 2015). Annotators focused on a single speaker at a time, but were able to listen to their utterances in context. Following the guidelines in Wilson (2008), they were instructed to consider not only direct expressions of opinion, but also agreement and disagreement, suggestion, arguing for or against something, stance soliciting questions and their responses, and beliefs from which sentiments can be inferred for stance annotation. Textual examples were provided to help them assign a stance strength to each spurt, but ultimately it was the annotator's perception that determined which level to use. Annotations

were reviewed by a second annotator and any inter-annotator disagreement resolved. Stance strength annotation guidelines are given in Table 3.1.

Label	Description	
0	No stance:	list reading, back channels
1	Weak stance:	cursory agreement, suggestions, bland opinion
2	Moderate stance:	stronger opinion, disagreement, alternatives, questioning
3	Strong stance:	very emphatic versions of #1 and 2
X	unclear:	stance cannot be determined

Table 3.1: Stance Strength Annotation Guidelines

Any spurt annotated with a stance strength label other than 0 was then annotated for polarity (positive, negative, neutral) using the same procedure.

3.3.1 PSI Corpus

The ATAROS tasks were, by design, low-stakes tasks designed to mitigate any effect of strong emotion. As a result, there is the potential that any results found in this data applies only to stance taking in a low-stakes, cooperative laboratory environment. To validate that these results were not a remnant of the controlled environment, nor that these behaviours manifest only in a collaborative setting, the ATAROS corpus includes data from an inherently adversarial, high-stakes, real world dialogic interaction was also annotated with stance strength. This data was from the US Senate Homeland Security Permanent Subcommittee on Investigations (PSI) regarding the 2007 – 2008 financial crisis. The specific portion of the hearing that was analyzed is the testimony and subsequent questioning of Lloyd Blankfein, the Chairman and CEO of Goldman Sachs, on April 27, 2010. Subcommittee members involved in the interaction are: Chairman Carl Levin (D-MI), John McCain (R-AZ), Ted Kaufman (D-DE), Tom Coburn (R-OK), Claire McCaskill (D-MO), and Mark Pryor (D-AR).

For this data, henceforth referred to as **PSI**, the audio recording was retrieved from C-SPAN,[1] and the transcription from the Federal Record[2] was used as a baseline. This transcription was verified against the recording. This involved delimiting the transcription into spurts in a TextGrid and adding disfluencies and filled pauses.

All subsequent alignment and annotation proceeded as it had with the ATAROS data.

Each senator's questioning of Lloyd Blankfein is considered a separate dyad for the purposes of this study. Since Lloyd Blankfein is common between all dyads, it allows for an investigation into how a single speaker's behaviour changes with different speaking partners. Additionally, unlike the ATAROS tasks which are collaborative, these interactions are combative, allowing us to show that any results are not due to a conscious effort on the part of the speakers to build a positive rapport with each other.

3.4 Data Preparation for Experiments

The data for this study was extracted from the TextGrids, described in sections 3.2 and 3.3. The text for each spurt was extracted and spurts were mapped to their speaker and stance annotation; where the annotation shows overlapping speech, the start time was used to determine order. In this way, the turn-taking nature of the dialogue is preserved. The text was cleaned of any voice quality annotations and valence markers (+ or -). Any spurts from which all of the textual content was removed (i.e. the speaking turn had only a voice quality annotation) were excluded. Spurts lacking stance annotation, or having been annotated with an "X" or "x" were examined. I decided to include them as utterances since they are still relevant to the dialogical nature of this study.

Table 3.2 shows the distribution of stance annotations for each corpus. ATAROS, as a laboratory-based collaborative task, is heavily centered around weak stance, while PSI, being a real-life combative situation centering around a serious topic is centered around

[1]http://c-span.org

[2]https://www.govinfo.gov/content/pkg/CHRG-111shrg57322/html/CHRG-111shrg57322.htm

strong stance. Stance strength annotation 3 made up less than 1% of each corpus, and was therefore combined with stance strength annotation 2. This is consistent with other projects involving a strength annotation (Wilson and Wiebe, 2003). It should be noted that the nature of the dialogues in each data set is inherently different. The ATAROS tasks are turn-based interactions with shorter speaking turns. The PSI corpus, on the other hand, includes prepared statements and questions; speaking turns are overall longer and less spontaneous.

Stance	**ATAROS 3I**	**ATAROS 6B**	**PSI**
0	2032 (22.9%)	2063 (25.7%)	656 (26.9%)
1	5520 (66.3%)	4117 (51.3%)	764 (31.3%)
2	1310 (14.8%)	1845 (23.0%)	1022 (41.9%)
Total:	8862	8025	2442

Table 3.2: Spurt Count per Stance Annotation

Because of the nature of the task in each corpus is so different, these corpora allow us to see stance taking in varied contexts. In the ATAROS 3I corpus, in addition to discussing the placement of the cards, there was discussion about how to conceptualize the task items, such as in the following. Note that the separation of a single speaker's speaking turn into spurts is shown by the stance strength annotation in parentheses at the end of each spurt.

A: Cake mix.. is a - (1) Introduction to baked goods. (1) We could put those things together. (2)

B: So the - (x) So the, like - (x) Not finished products could go between it and sugar. Is that what you're thinking? .. I mean not - Not these cuz those are made and those are made, or what are you thinking? (2)

A: Pa. Well, the three of these things are made, so packages of cookies, boxes of donuts, and bagels. (2)

B: These are made. (1) So these are already baked. (1)

Among the stance strength annotations in this example, note the mix of 1's and 2's. These were likely annotated thusly on the basis of the acoustics of the utterances rather than the textual content since there is nothing in the content of the spurts that signals that some are stronger than others. For those spurts marked with the (x), multiple annotators agreed that it was not clear which level of stance was being expressed. Among the spurts annotated with stance strength label 0, many of them were of a confirmatory nature, such as an echoing of task items, or relayed a status of the task to date such as in the following:

A: We're done. (0)

B: Okay. (0) We got a nice store. (1)

A: Yeah. (0)

B: People won't have trouble finding things in this one. (1)

The ATAROS 6B corpus, on the other hand, revolved around more civically oriented issues. Speakers were more prone to discussing the issues, and draw from their life experience than they were in the 3I task.

A: Mm-kay. .. Reusable bag campaign. (1) Uh do we really need to campaign over it anymore? (2)

B: It's pretty - (1) It's - I know. People are doing pretty good. (2)

A: People are doing well and for the most part their keeping their plastic bags anyway. And - (2)

B: Yeah. (1) At least - at least now, in here, if we're doing that. (1)

A: Well - I'm gonna assume that we're - .. in the perfect society that has now adopted the reusable bag campaign. (2)

B: Yeah. Yeah. Me too. (1)

In addition to the tendencies listed above, many of the spurts annotated with stance label 0 acted as floor holders, often accompanied by a pause:

A: Yeah, what else would you cut? (1)

B: [*4.1 second pause*] What else would I cut? (0)

A: Other than - Yeah. You're gonna cut .. sugar-free juice and - (0)

Finally, the PSI corpus was a formal hearing. Many of the stance-less spurts were part of the senatorial decorum, the speaking turns were longer, and the nature of the setting lended itself to longer, more coherent thoughts. There was often a performative aspect apparent in the speaking style of the subcommittee members, and Lloyd Blankfein tended to be very careful in choosing his words, so he frequently showed hesitation. Both of these tendencies are shown in the following:

PRYOR: Thank you, Mister Chairman. (0) Uh, I'd like to start, if I can, (1) about - (0) with the (0) topic of asking you about (0) um .. credit rating agencies. In retrospect, .. uh how accurate .. were the credit rating agencies (0) in rating the various tranches of CDOs? (0)

BLANKFEIN: Uh, in retrospect, um .. they were .. uh inaccurate. (1)

The different nature of these corpora will be useful in that any findings from the experiments will be applicable across domains and dialogical settings. The ATAROS 3I task scenario is on the more artificial side, while one could imagine discussing civic issues such as the ones that came up in the ATAROS 6B task. Both of these are cooperative and friendly dialogues. The PSI corpus, on the other hand, is a real-word instance of stance taking with high stakes in an inherently adversarial setting.

3.5 Dialogical Stance Trends

As a preliminary investigation, the first test is to establish whether one speaking partner's expressions of stance have any influence on the other partner. Rather than each spurt, the unit of measurement here is the speaking turn. A speaking turn is comprised of sequential spurts uninterrupted by the other speaker. Where there were multiple spurts within a speaking turn, the highest stance strength annotation of the component spurts was assigned to the speaking turn as a whole.

Figure 3.3 shows that there is a conditioning effect of the speaking partner's use of stance. By stance, I am counting any speaking turn annotated with a 1 or above. Had there been no effect at all, the conditional and the marginal probabilities would have been equal for every speaker. I have made the point size for the higher probability, marginal or conditional, larger so that it stands out. The data shows that the conditional probability is greater than the marginal for over 70% of speakers in both tasks.

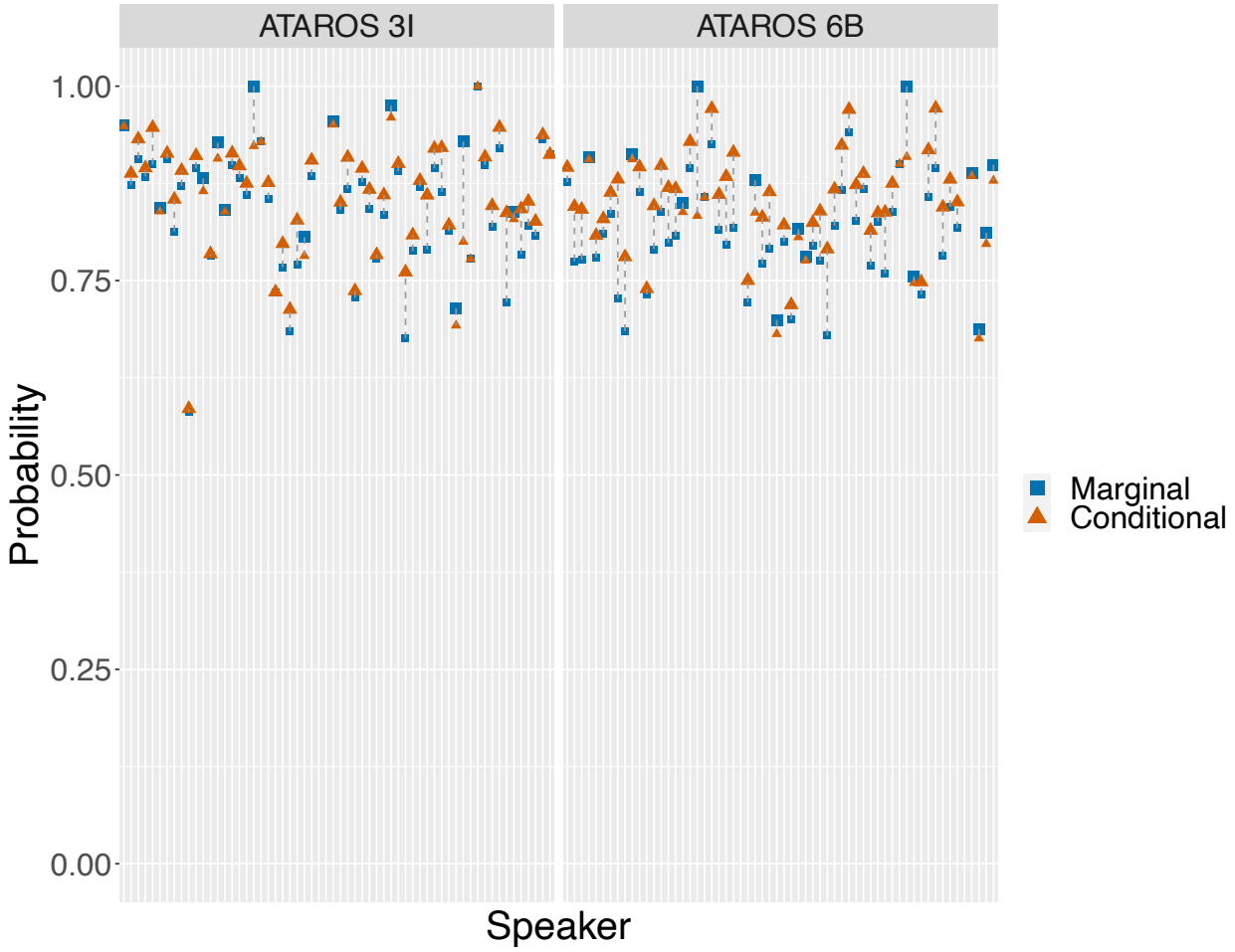


Figure 3.3: P(Stance) and P(Stance|Previous Stance) in the ATAROS Corpora
Point Size Indicates the Greater of the Two Probabilities

Figure 3.4 shows the same comparison for strong stance, strength annotation 2. Here, again the conditioning effect of the speaking partner using strong stance in the immediately preceding speaking turn is clear. As with the previous figure, the larger point size indicates the higher probability, which again is the conditional probability for over 70% of speakers.

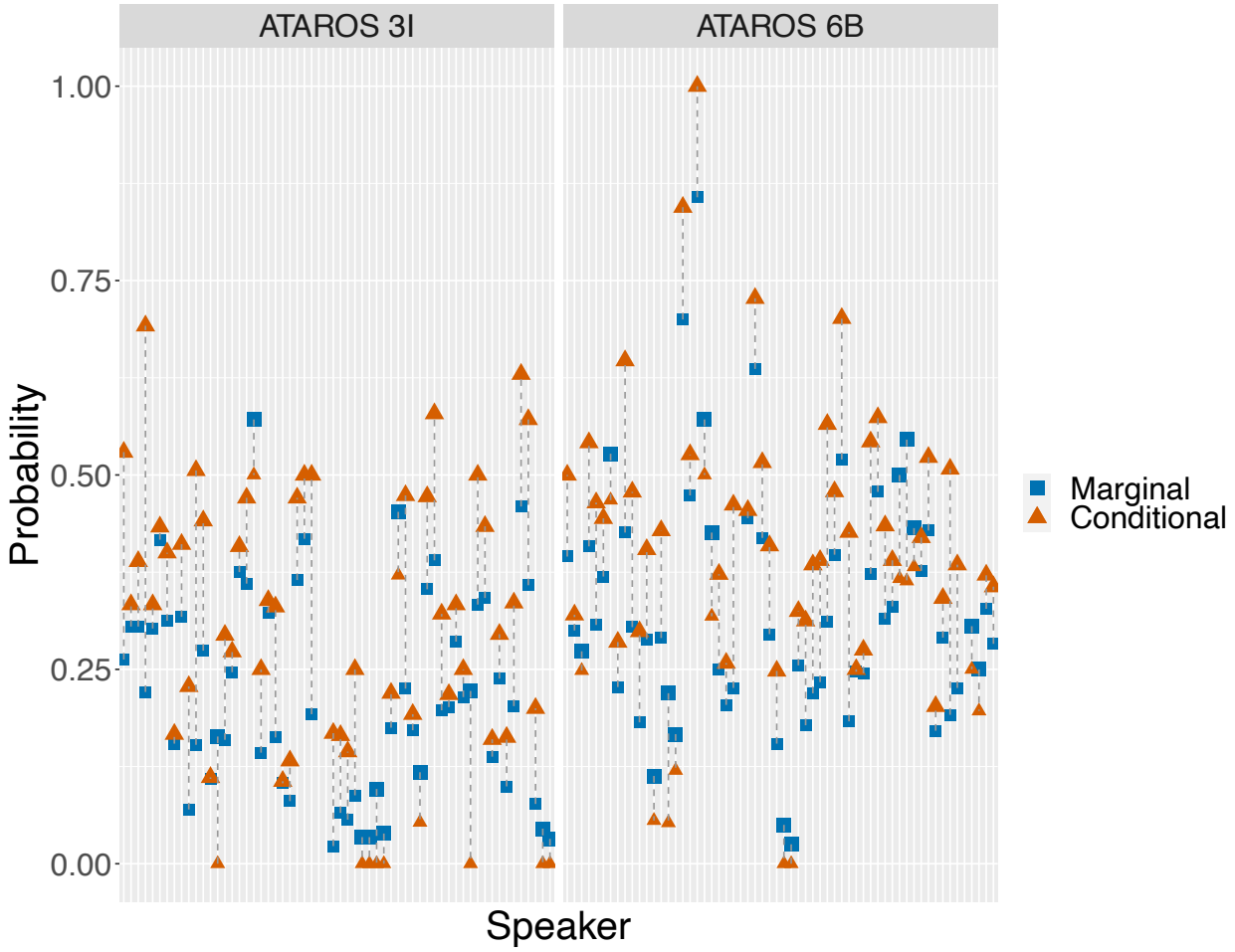
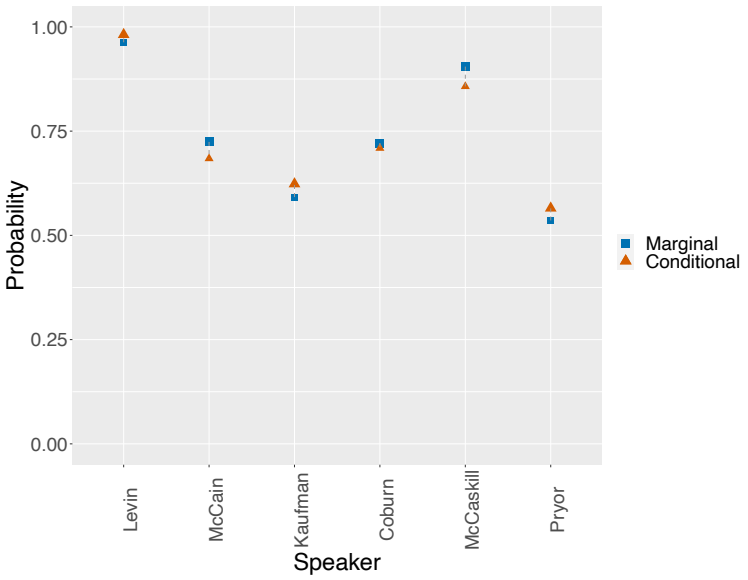
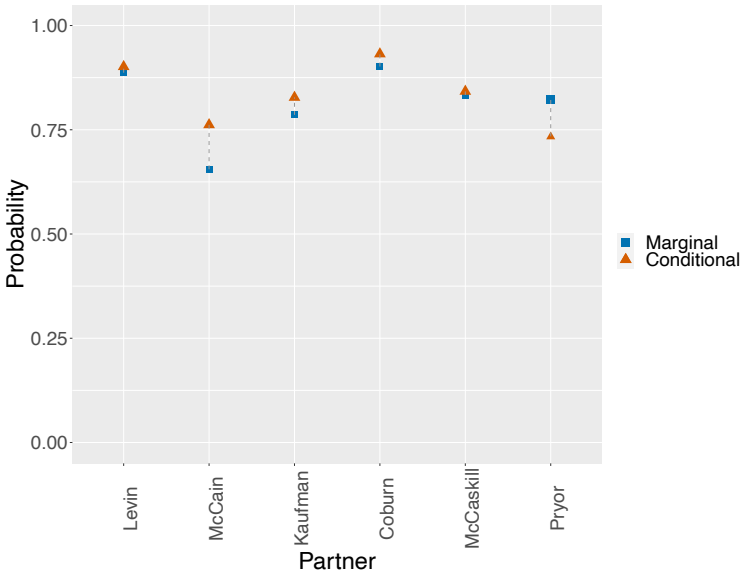


Figure 3.4: P(Strong) and P(Strong|Previous Strong) in the ATAROS Corpora
Point Size Indicates the Greater of the Two Probabilities

Figure 3.5 shows the conditioning effect of stance for the PSI corpus. Subfigure 3.5a shows the Subcommittee members during their questioning of Lloyd Blankfein, while Subfigure 3.5b shows Lloyd Blankfein in his responses.



(a) Subcommittee Members



(b) Lloyd Blankfein

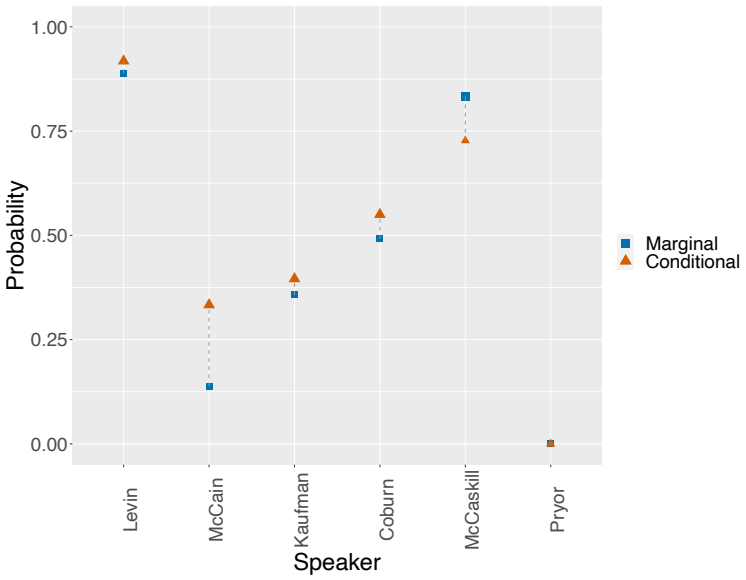
Figure 3.5: P(Stance) and P(Stance|Previous Stance) in the PSI Corpus

Point Size Indicates the Greater of the Two Probabilities

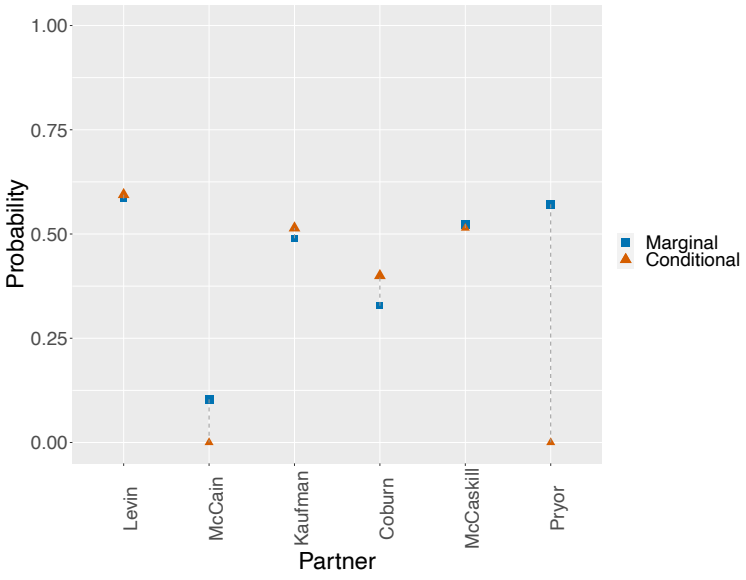
Subfigure 3.5a shows that Lloyd Blankfein's use of stance did not have much of a conditioning effect on the subcommittee members. This is not surprising given the fact that there was a power differential between him and the members of the subcommittee; those in power have an influence on the linguistic behaviour of those less powerful rather than the other way around (Danescu-Niculescu-Mizil et al., 2012). This effect is shown in Subfigure 3.5b where for five of the six speaking partners, their expression of stance increased Blankfein's use of stance.

Figure 3.6 shows the conditional and marginal probability of strong stance. Subfigure 3.6a shows the Subcommittee members, and Subfigure 3.6b shows Lloyd Blankfein in his responses.

Here, there is a conditioning effect, though not a large one, for the majority of subcommittee members, and for Lloyd Blankfein in his interactions with them. Subfigure 3.6a shows that Lloyd Blankfein's use of strong stance increases the probability of strong stance for all subcommittee members except Claire McCaskill and Mark Pryor. Mark Pryor is notable in this corpus for not having used strong stance at all; this is reflected in Subfigure 3.6b, where the conditional probability of Lloyd Blankfein using strong stance in his interaction with Pryor is 0. Otherwise, Carl Levin and Claire McCaskill look to have had very little to no conditioning effect, while Ted Kaufman and Tom Coburn's use of strong stance conditions Lloyd Blankfein to use strong stance in his responses and John McCain's makes him less likely.



(a) Subcommittee Members



(b) Lloyd Blankfein

Figure 3.6: P(Strong) and P(Strong|Previous Strong) in the PSI Corpus

Point Size Indicates the Greater of the Two Probabilities

Based on these measures, there is evidence that the use of stance by one speaker influences the use of stance by the other speaker, and even stronger evidence that strong stance conditions strong stance in the other speaker. This effect is stronger in the co-operative interaction (ATAROS) than in the adversarial interaction (PSI). Reasons for this could the desire to build and demonstate rapport with their speaking partner in the ATAROS corpus, which was unnecessary in the PSI corpus due to the adversarial nature of the interaction and the formality of the environment in which it occurred. Additionally, speakers entered into the PSI interactions with a pre-established view of the issue they intended to argue; their goal of the interaction was to express this, which by its very nature involves stance-taking unprompted by the speaking partner.

The remainder of this dissertation will investigate specific lexical and grammatical choices speakers demonstrate when they are expressing stance in a dialogical setting versus when they are not.

Chapter 4
EVALUATIVE STRENGTH OF LEXICON

Psychological studies have analyzed words along three subjective dimensions, *valence*, the negative to positive dimension; *arousal*, the passive to active dimension; and *dominance*, the submissive to dominant dimension (Russell, 1980). These dimensions work largely independently of each other. Given a set of nearly synonymous words, a speaker will choose one that best matches their preferences along one or all of these dimensions.

Speakers are conscious of their word choices. While not all word choices require careful thought, it is not difficult to recall a time when we had to carefully choose our words, most frequently to uphold politeness; additionally, when we have mis-chosen a word, we recognize it instantly. Often when we are selecting a word, our alternatives vary in strength. Compare alternatives such as *dislike* and *hate* or *love* and *like*.

This chapter addresses the use of word strength along the dimensions of *valence*, *arousal*, and *dominance*, and the relationship of these dimensional scores to stance stength. I predict that strength along all three dimensions will be perceived as strong stance. Additionally, since this is a dialogical study, I investigate whether one speaking partner's use of strong language along these dimensions influences the strength of their speaking partner's word along the same dimension.

4.1 NRC-VAD Lexicon

The word strength scores used in this study are from the the **National Research Council of Canada Valence, Arousal, and Dominance Lexicon** (NRC-VAD Lexicon), a scored lexicon containing over 20 000 emotive words. The words were manually curated from previous affect lexicons, supplemented with other external sources (Mohammad, 2018). Words

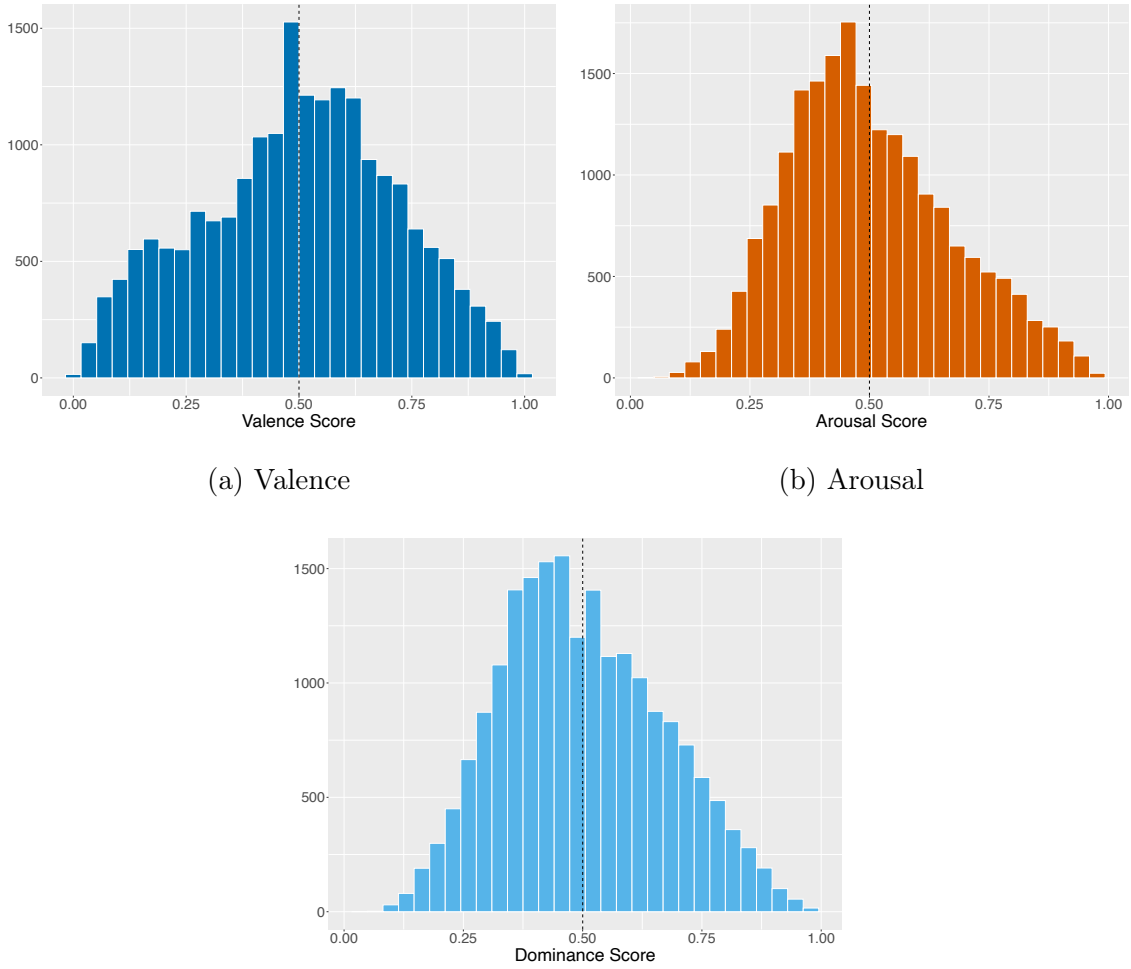
were scored along the dimensions of valence, arousal, and dominance through crowdsourcing and scores were aggregated using a method called Best-Worst Scaling (Louviere and Woodworth, 1991). Best-Worst Scaling presents annotators with a 4-tuple of words where they are asked to select the word that most strongly and least evokes the specific dimension to them, using the words given in Table 4.1 as examples of the two ends of the dimensional range. Words were arranged into 4-tuples in such a way that each word is seen in eight 4-tuples, and no 4-tuples have more than two words in common. Annotators were tasked with annotating a single dimension. An aggregate score for each word, along each dimension, was calculated by subtracting the proportion of times a word was selected as the least evocative of the dimension from the proportion of times a word was selected as most strongly evoking the dimension. This was then linearly transformed to a [0 - 1] scale where 0 represents the low end of the scale (negative/not aroused/submissive) and 1 represents the high end of the scale (positive/aroused/dominant). Note that most of the words given in Table 4.1 come in oppositional pairs. This indicates that a score of 0.5 represents neutrality (neither positive nor negative/neither aroused or not aroused/neither dominant nor submissive).

Dimension	Low	High
Valence	unhappiness	happiness
	annoyance	pleasure
	negativeness	positiveness
	dissatisfaction	satisfaction
	melancholy	contentedness
	despair	hopefulness
Arousal	unarousal	arousal
	passiveness	activeness
	relaxation	stimulation
	calmness	frenzy
	sluggishness	jitteriness
	sleepiness	alertness
	dullness	
Dominance	submissive	dominant
	controlled by outside factors	in control of the situation
	weak	powerful
	influenced	influential
	guided	autonomous
	cared for	
		important

Table 4.1: Descriptive Words for each Dimension Used by Mohammad (2018)

Figure 4.1 shows the distribution of scores along each dimension for the entire NRC-VAD Lexicon. Note that the majority of the scores fall in the vicinity of the neutral point for all three dimensions. The valence scores show a slight skew toward the positive end of the range, while the arousal and dominance scores show a slight skew toward the lower end of

their ranges (unexcited and submissive). Recall from Table 4.1 that not all terms in the lower end of the arousal and dominance scales carry negative connotation; *relaxation* and *calmness* connote positive emotions, with valence scores of 0.873 and 0.934 respectively.



(a) Valence

(b) Arousal

(c) Dominance

Figure 4.1: Distribution of Scores in the NRC-VAD Lexicon

For this study, I am using only the words from the NRC-VAD Lexicon that appear in the ATAROS and PSI corpora.

Text from each spurt was first case normalized then tokenized using the NLTK function `word_tokenize` (Loper and Bird, 2004). Disfluencies, that is words that were started yet not finished, were removed.[1] Lemmatization was not necessary since the NRC-VAD Lexicon has scores for different inflected forms. Task vocabulary, which is given in Appendix A, was removed prior to querying the NRC-VAD lexicon so that any strongly emotive task-related terms, used in these contexts in a non-emotive sense, did not falsely inflate the emotional strength of the utterance. For the ATAROS tasks, the task vocabulary was extracted from Freeman (2015), then manually expanded to include other terms such as synonymous, alternate, and shortened versions of the vocabulary items. For the PSI corpus, the only topic-related term that held any strongly emotive sense was *crisis*, which was used invariably to refer to the event, the financial crisis. This was, therefore, treated as task vocabulary for the purposes of this experiment.

Table 4.2 shows the count of words from the NRC-VAD Lexicon that was used in each corpus.

Corpus	# **Emotive Tokens**
ATAROS 3I	974
ATAROS 6B	1499
PSI	1255

Table 4.2: Number of Words from the NRC-VAD Lexicon Appearing in the Data

Note that the ATAROS 3I and 6B corpora included the same dyads. The difference in the count of emotive words is likely due to the stronger engagement evoked by the 6B task.

Figure 4.2 shows the strongest scoring words along each dimension for all three corpora. Dimensional strength is represented by the magnitude of the distance of the score from the neutral point. That is, scores close to the two ends of the ranges (0 and 1) are strong along

[1]See Chapter 3 for annotation guidelines

that dimension, while scores around 0.5 are not. In these word clouds, dimensional strength is represented by font size; the larger the font, the stronger the word.

(a) ATAROS 3I: Valence (b) ATAROS 3I: Arousal (c) ATAROS 3I: Dominance

(d) ATAROS 6B: Valence (e) ATAROS 6B: Arousal (f) ATAROS 6B: Dominance

(g) PSI: Valence (h) PSI: Arousal (i) PSI: Dominance

Figure 4.2: Strongest Scoring Words along each Dimension

Strength is Represented by the Magnitude of the Distance from the Neutral Point

Table 4.3 shows the count of words from the NRC-VAD Lexicon shared across the three corpora.

	3I	6B	PSI
3I		496	371
6B			537

Table 4.3: Count of Overlapping NRC-VAD Lexicon Words

This shows the variety of emotive vocabulary used. Only half of the emotive words used in the ATAROS 3I task also appeared in the ATAROS 6B task, in which the same speakers were partnered with the same partners; both tasks were completed during the same hour-long session. Given that the ATAROS 6B task has a similar number of emotive words as the PSI corpus (1499 and 1255 respectively) it is also noteworthy that only 537 words are used in both.

Table 4.4 shows the count of spurts containing a word from the NRC-VAD Lexicon. Note that the high spurt count in the ATAROS 3I corpus is due to there being more, shorter spurts, rather than it being a longer conversation than the ATAROS 6B task. Similarly, in the PSI corpus the speaking turns were longer overall.

Corpus	**Total Spurts**	**Spurts with Emotive Word**
ATAROS 3I	7805	4529 (58%)
ATAROS 6B	7025	4431 (63%)
PSI	2334	1969 (84%)

Table 4.4: Count of Spurts per Corpus

From here it is clear that the PSI corpus contains the highest concentration of emotive words, while the ATAROS 3I corpus contains the lowest. Recall from Table 4.2 that the ATAROS 6B task has considerably more emotive words than the ATAROS 3I task, yet the raw count of spurts containing an emotive word differs only by a small amount. This is an

indication that, not only is the proportion of emotive spurts in the ATAROS 6B task greater, spurts likely contains multiple emotive words.

4.1.1 Valence Scores

The term *valence* refers to the positive $\leftrightarrow$ negative dimension of a word, where 0.0 represents the peak of negative emotion, and 1.0 represents the peak of positive emotion; 0.5 represents the neutral point, neither positive nor negative.

Figure 4.3 shows the distribution of valence scores for words used in the ATAROS and PSI corpora respectively. As in the NRC-VAD Lexicon itself, there is concentration at the neutral point (0.5) and a skew toward the positive end of the range (> 0.5). The difference in the y-values between these corpora represents the fact that there were more emotive words used in the ATAROS 6B and PSI tasks than the ATAROS 3I task.

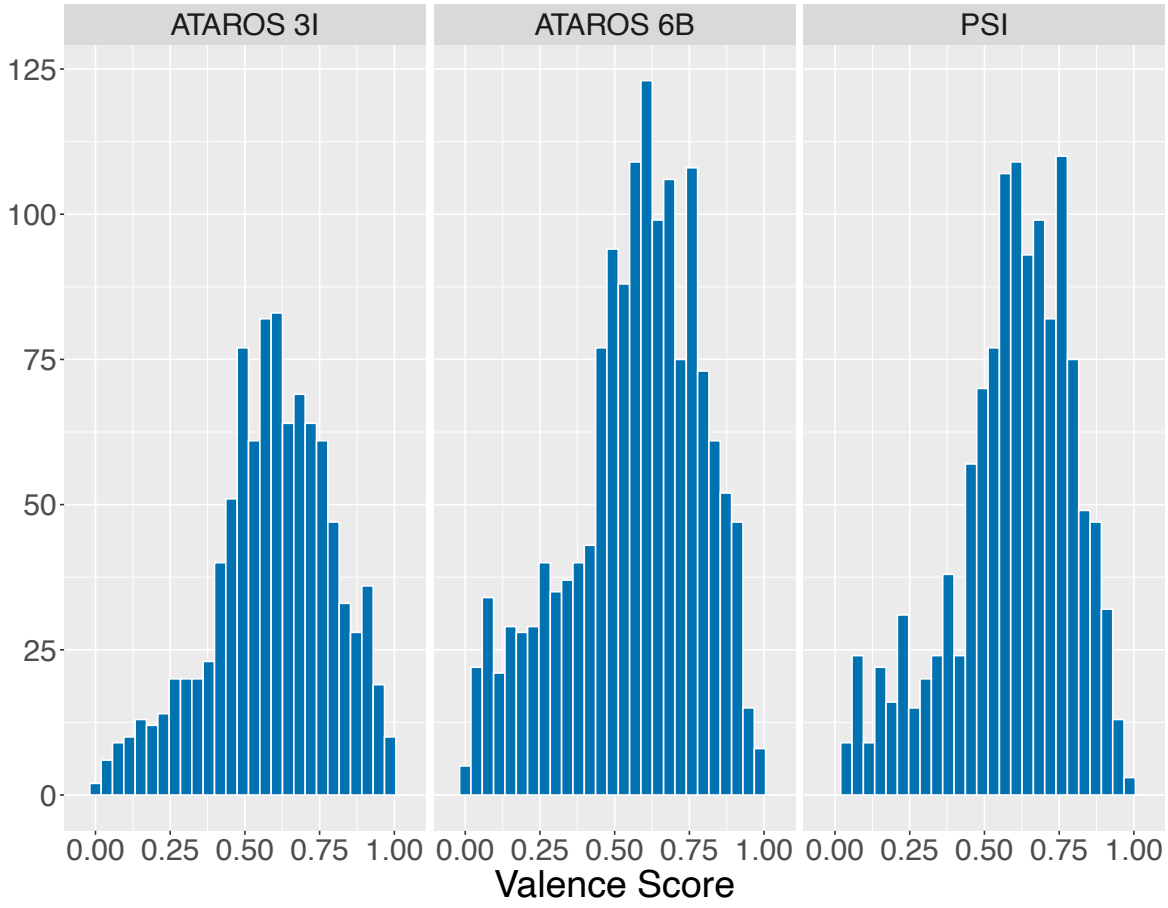


Figure 4.3: Distribution of Valence Scores Among Words Used in the Corpora

That there is more weight in the positive valence ranges speaks to the tendency to keep the tenor of the dialogues positive (Lakoff, 1973). Even in the inherently contentious PSI corpus, the tendency to use positive words is apparent. Even when being critical or disagreeing, it is more likely to be framed using positive words rather than overly negative words (Lakoff, 1973).

4.1.2 Arousal Scores

Arousal reflects the amount of excitement expressed by a word (Mohammad, 2018). It is represented on a bipolar [0, 1] scale where 0 represents a complete lack of excitement, and 1 represents extreme excitement.

Figure 4.4 shows the distribution of arousal scores in the three corpora. All three skew toward the "unexcited" end of the score range.

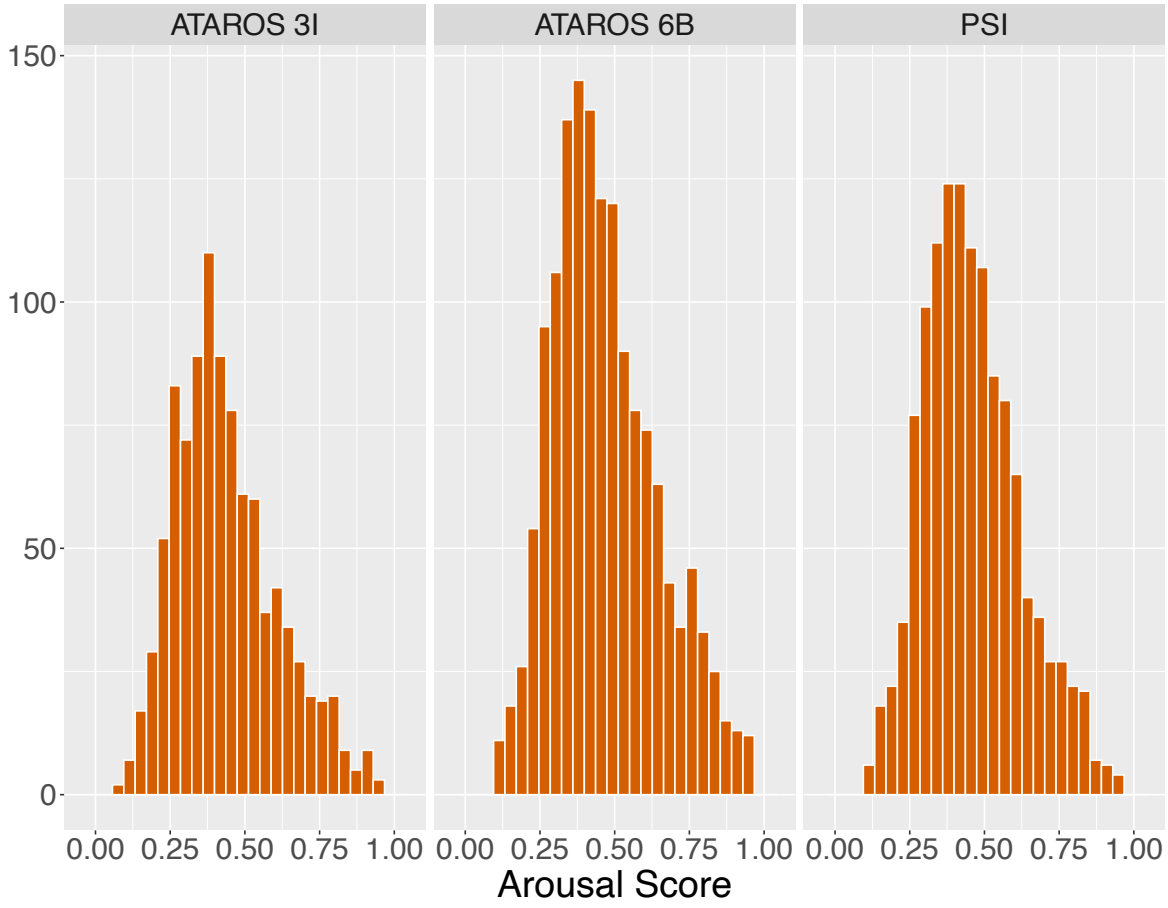


Figure 4.4: Distribution of Arousal Scores Among Words Used in the Corpora

Recall that these emotional dimensions are largely independent of each other. Figure 4.5 shows a mapping of words appearing in all three corpora along the dimensions of *valence* (x-axis) and *arousal* (y-axis). Lines at the 0.5 mark of both axes represent neutrality along the respective dimension. Note that other than there being greater weight in the positive-valence, low-arousal quadrant, there is no discernible pattern in the data.

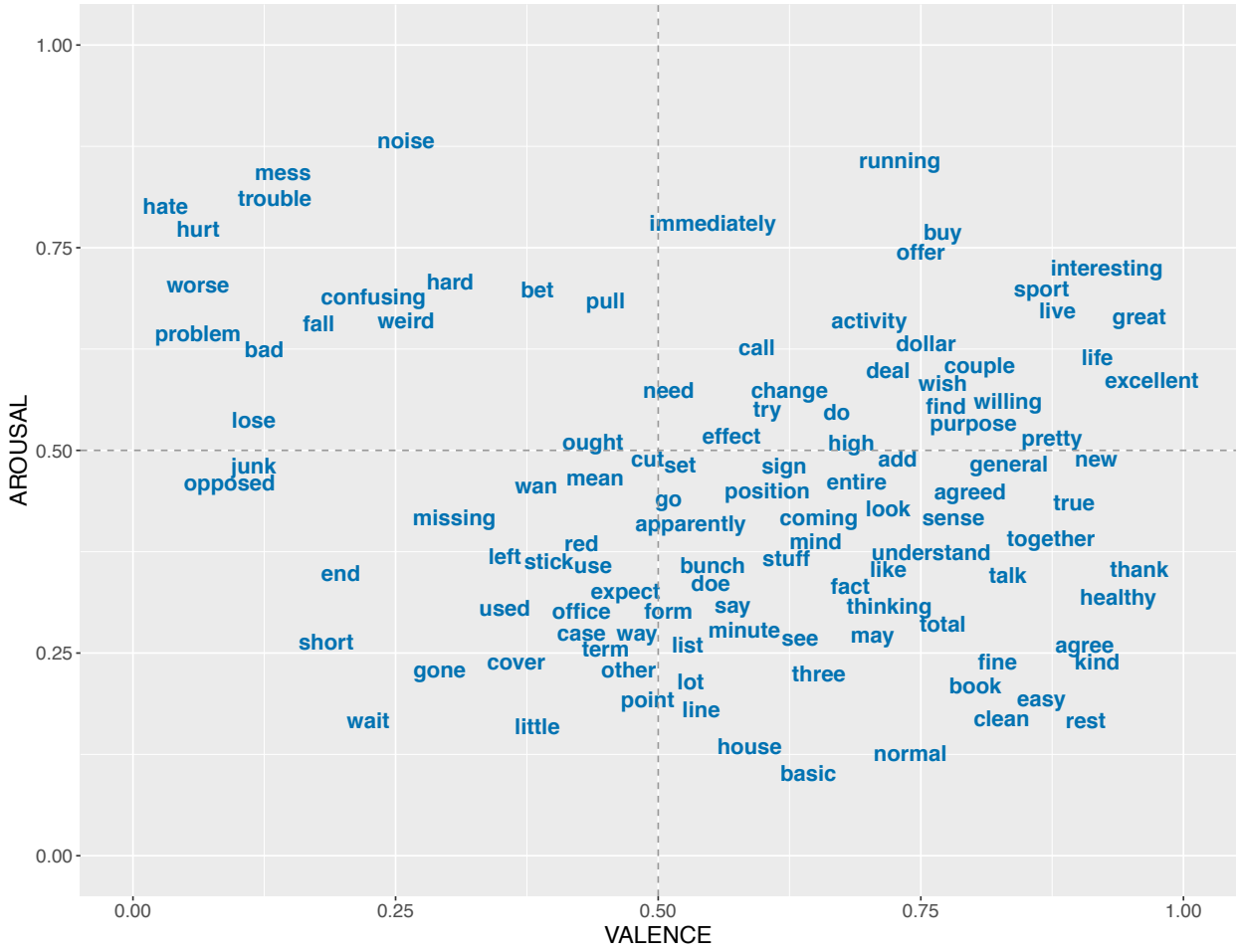


Figure 4.5: Scores Along the Dimensions of Valence and Arousal

4.1.3 Dominance Scores

The term *dominance* refers to the dominant ↔ submissive or *in control* ↔ *controlled* dimension of a word. It is represented on a [0,1] scale where 0 represents peak submissiveness end and 1 represents peak dominance.

Figure 4.6 shows the distribution of dominance scores in the three corpora. The ATAROS 3I corpus is skewed toward the submissive end; ATAROS 6B is centered at the neutral point. This speaks to the cooperative nature of these tasks. The PSI corpus, on the other hand, shows more weight in the dominant end of the range, which is not surprising given

the speakers and the situation in which the dialogue is captured; one would expect that political and industry leaders would use more dominant language in general, particularly in the context of a hearing.

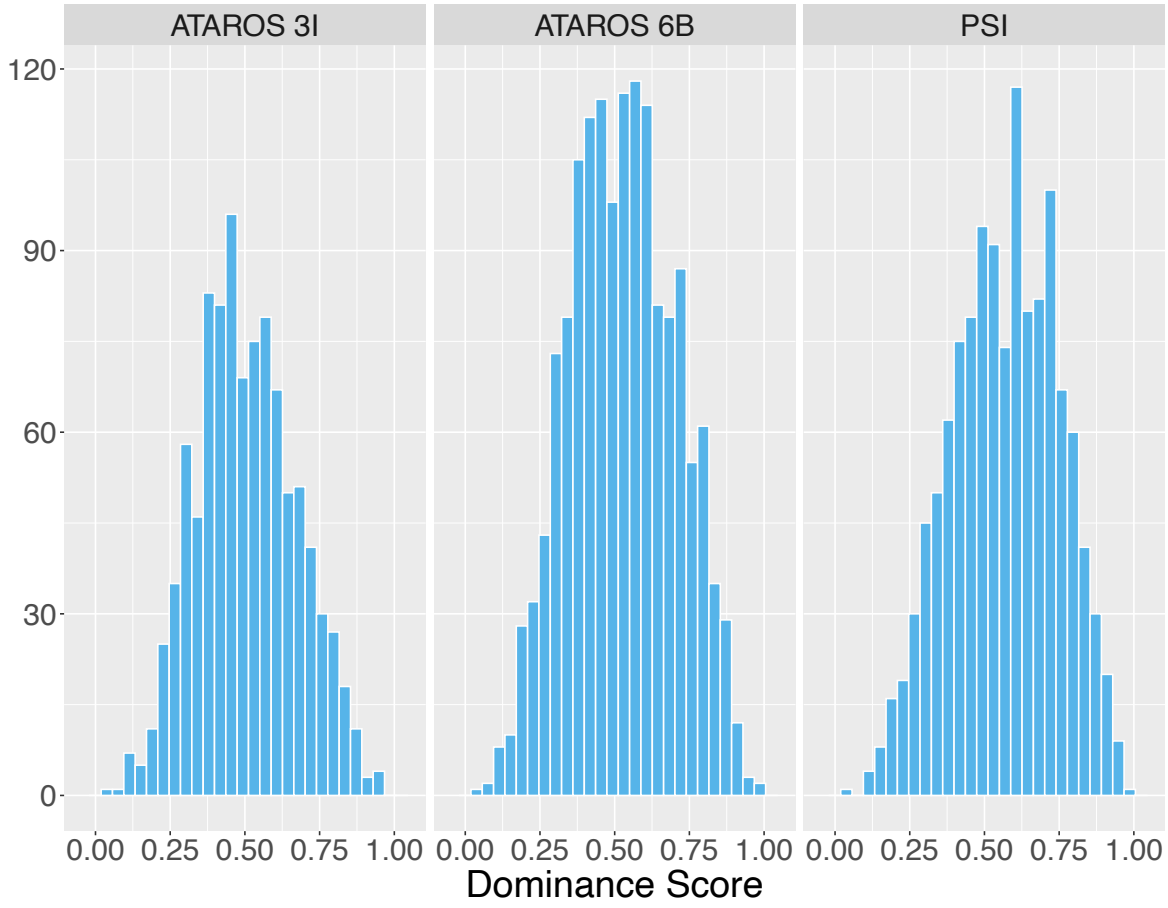


Figure 4.6: Distribution of Dominance Scores Among Words Used in the Corpora

Figure 4.7 shows a mapping of words appearing in all three corpora along the dimensions of *valence* (x-axis) and *dominance* (y-axis). Lines at the 0.5 mark of both axes represent neutrality along the respective dimension.

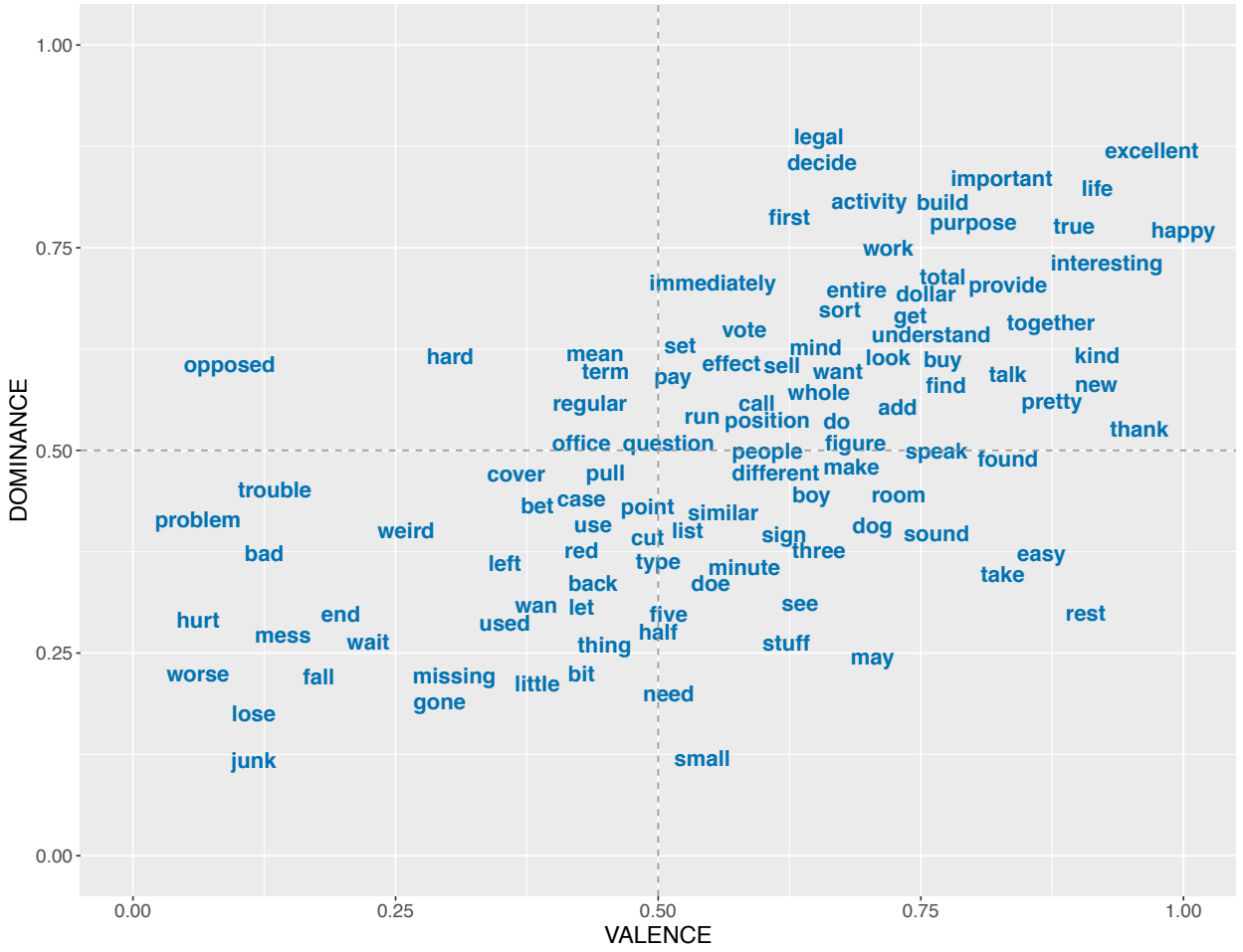


Figure 4.7: Scores Along the Dimensions of Valence and Dominance

Along these dimensions, one could picture a fitted diagonal line roughly corresponding to y=x of the Cartesian plane, indicating some relationship between valence and dominance: dominance tends to align with positive valence, and submissiveness tends to align with negative valence.

4.2 Experimental Design

Some experiments use the spurt as the unit of measurement, and others use the speaking turn. This will be specified for each experiment. Since the unit of annotation is the spurt, sequential spurts by a single speaker were combined to form a speaking turn. The stance strength annotation assigned to the speaking turn was the highest strength annotation among the component spurts.

For each spurt and speaking turn the stance strength annotation (0, 1, or 2), the number of NRC-VAD Lexicon words, and the scores for these words along the dimensions of valence, arousal, and dominance, were recorded.

4.2.1 NRC-VAD Lexicon Word Count and Stance

First, to establish whether there is a relationship between words from the NRC-VAD Lexicon and stance, a series of Pearson Chi-squared (χ^2) tests (Pearson, 1900) were run. The first test shows that the distribution of spurts containing words from the NRC-VAD Lexicon among stance-laden (stance strength labels 1 and 2) and stance-less (stance strength label 0) spurts is statistically significant ($p < 0.001$) for all three corpora. The second test shows the distribution of NRC-VAD Lexicon words also varies significantly with stance strength label ($p < 0.001$) for all three corpora.

From this we can conclude that there is a relationship between the use of words from the NRC-VAD Lexicon and stance strength. Any conclusions regarding a spurt-level count of NRC-VAD Lexicon terms and stance strength is confounded by the fact that spurts expressing stronger stance tend to be significantly longer ($p < 0.001$). Therefore, a proportional measurement would not serve to mitigate these effects.

4.2.2 NRC-VAD Lexicon Scores and Stance

As for the scores assigned to words along the dimensions of *valence*, *arousal*, and *dominance* themselves, to test whether these have any relationship to stance strength, a multinomial logistic regression model was built using the R package `nnet` (Venables and Ripley, 2002) for each corpus and dimension combination, for a total of 9 models. These models use stance strength as the dependent variable, and the relevant dimensional score as the independent variable. Since many spurts contain multiple words from the NRC-VAD Lexicon, I assigned the maximum score along each dimension as the spurt-level dimensional score. A Log Liklelihood Ratio Test, using the R package `lmtest` (Zeileis and Hothorn, 2002) was used to determine statistical significance of these scores. This test compares the model with the respective dimension scores against a saturated model, that is, a model with no independent variables. All dimensions showed statistical significance ($p < 0.001$) for all corpora.

4.2.3 Corpus Breakdown

Because subjectivity, and therefore stance, is interpretable only in context (Wiebe and Mihalcea, 2006), to truly determine the effect of word strength on stance strength, the unit of measurement should be on the spurt or speaking turn level. The breakdown of stance strength labels in the three corpora is given in Table 4.5.

	Spurts			**Speaking Turns**		
	0	**1**	**2**	**0**	**1**	**2**
ATAROS 3I	2032	4463	1310	971	3224	1161
ATAROS 6B	2063	3117	1845	984	2090	1488
PSI	656	656	1022	151	130	292

Table 4.5: Distribution of Stance Labels among Spurts and Speaking Turns

Recall that a speaking turn is a series of sequential spurts by a single speaker. The stance strength assigned to the speaking turn is the highest stance strength annotation among the individual spurts.

4.2.4 Independent Variables

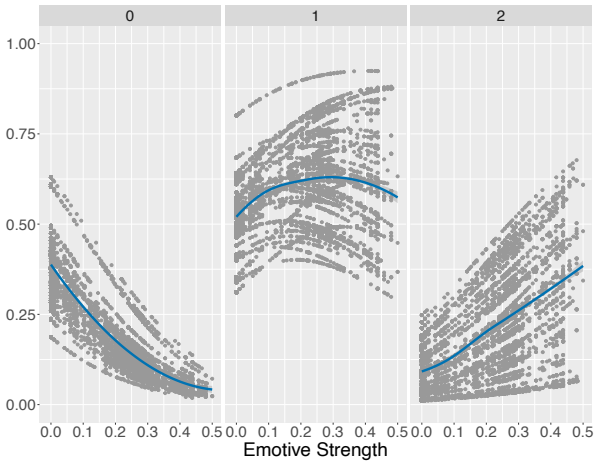
Thus far, the models have been treating each spurt or speaking turn as an independent event. This assumption is not accurate, since the focus of the each corpus is discourse between speaking partners, who inherently contribute multiple speaking turns. One would assume that individual speakers would vary with respect to individual tendencies along these dimensions. To verify that there is inter-speaker variation in word strength along these three dimensions, a set of multinomial logistic regression models, one per corpus, was trained, this time using speaker as the dependent variable and the respective dimensional score as the independent variable. The unit of measure here is the spurt. Again, when tested against a saturated model, all models showed statistical significance ($p < 0.001$) for all corpora. Therefore, to definitively test the effect of these scores, the speaker must be included in the model as a joint effect, and the effect of the dimensional scores should be tested in a Likelihood Ratio Test against a model with only *speaker* as a variable.

4.3 Results: Valence

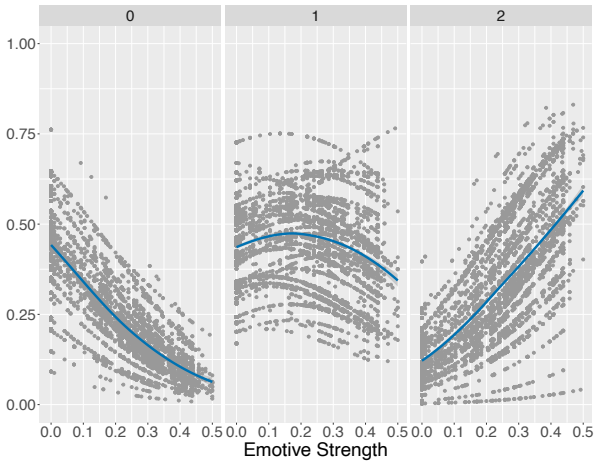
To reiterate from section 4.1.1, the term *valence* refers to the positive $\leftrightarrow$ negative dimension a word, where 0 represents the most negative emotion, and 1 represents the most positive emotion. 0.5 represents the neutral point, neither positive nor negative. As such, **emotive strength** is represented by the magnitude of the score from the neutral point in either direction, and is represented on a scale from [0.0 – 0.5].

This study seeks to determine whether the emotive strength of the words used have an influence on the perceived strength of the stance expressed. I say perceived because I am using the human-annotated stance strength labels used in the ATAROS corpus (Freeman, 2015). Going into the study, I predict that the strength of the words used affects the perceived strength of the stance being expressed.

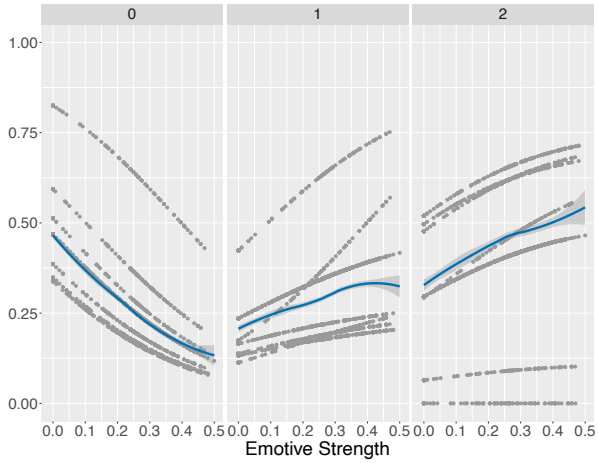
To test whether there is a relationship between emotive strength and stance strength, multinomial logistic regression models were trained using the `multinom` function from the R (R Core Team, 2015) library `nnet` (Venables and Ripley, 2002). One model was trained for each corpus. These models use **stance** as the dependent variable, and the **emotive strength** score (i.e. the magnitude of the distance from the neutral point) and **speaker** as joint independent variables. Against a model using only speaker as the independent variable, all three corpora show statistical significance ($p < 0.001$) in a Likelihood Ratio Test using the function `lrtest` from the R library `lmtest` (Zeileis and Hothorn, 2002). Figure 4.8 shows the fitted values generated from these models; they indicate how the probability of each stance strength changes as the emotive strength score increases. The emotive strength score is on the x-axis (a 0 - 0.5 scale), and the probability of the stance strength label is on the y-axis, for all three corpora. Lines were fitted using the LOWESS (Locally Weighted Scatterplot Smoothing) algorithm (Cleveland, 1981), a non-parametric, robust locally-weighted regression that uses weighted least squares to find the line that minimizes variance.



(a) ATAROS 3I



(b) ATAROS 6B



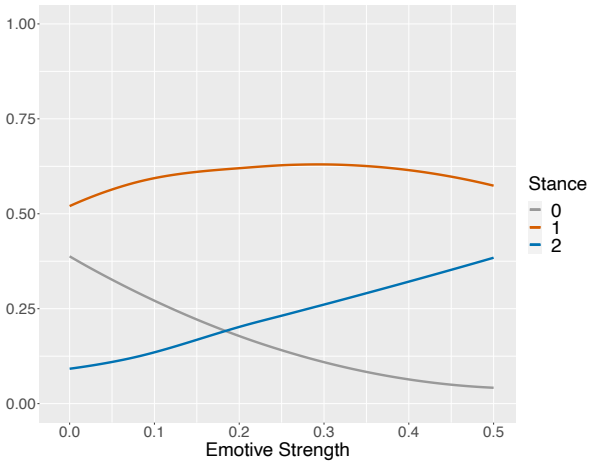
(c) PSI

Figure 4.8: Stance Strength as a Function of Emotive Strength

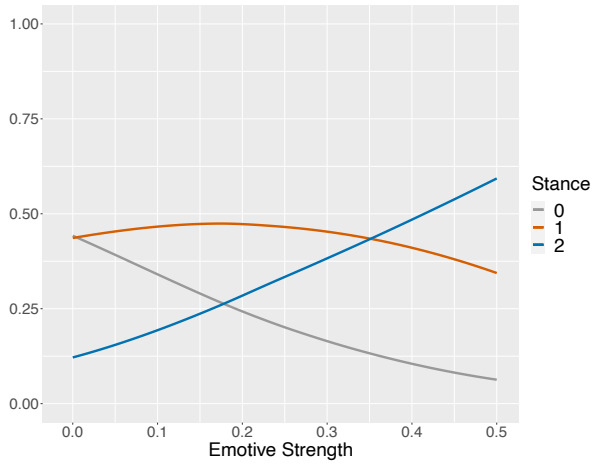
Looking at the fitted line, it is obvious that the probability of stance label 0 decreases and the probability of strong stance, stance label 2, increases as the emotive strength score increases for all three corpora. The trend is less clear for stance label 1 where the probability reaches a maximum point before decreasing for both ATAROS corpora. Note that these scales represent the distance from the neutral point, a [0 – 0.5] scale, rather than the entire

[0 – 1] valence scale, meaning that a score of 0.25 on these graphs represents a valence score of 0.75 (positive valence) or 0.25 (negative valence) on the original scale, not valence neutrality. This indicates that these maxima are at the mid-range of positivity or negativity. The PSI corpus, on the other hand, shows a monotonic increase throughout the score range.

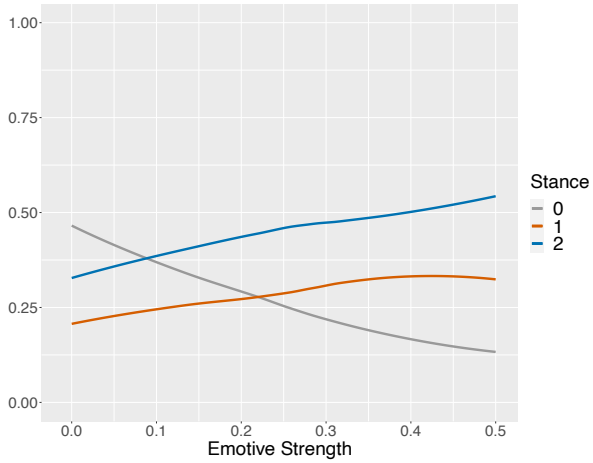
Figure 4.9 shows the fitted lines on a single plot, so that we can compare the stance strengths more easily. When you compare the stance strength annotations within a single corpus, it is clear that the most probable stance annotation for the ATAROS 3I task is 1, regardless of the emotive strength. In the ATAROS 6B corpus, stance strength 1 is the most probable at emotive strength scores below 0.35. Above this, stance strength 2 becomes the most probable; note that this represents valence scores at the extreme ends, < 0.15 and > 0.85 on the original scale. As for the PSI corpus, stance strength 2 is the most prominent score above scores of 0.1; on the original valence scale, scores below 0.1 represent the neutral range, (0.4 – 0.6).



(a) ATAROS 3I



(b) ATAROS 6B



(c) PSI

Figure 4.9: Stance Strength as a Function of Emotive Strength: Fitted Lines

From this, we can conclude that the emotive strength of the word definitely affects the perceived strength of the stance being expressed. Strong stance is expressed using strongly emotive words, at least in highly-engaged tasks. At lower levels of emotive strength, the probability of weak stance, stance label 1, increases as emotive score increases.

Now that I have shown that spurts containing a stronger emotive word are more likely to be perceived as expressing stronger stance, to show whether strongly emotive words are more likely to be used to express strong stance, the conditional probability of each stance strength label, conditioned on emotive strength, is compared to the prior probability of the stance strength. These measurements are made on the level of the speaking turn. The score assigned to each speaking turn for each dimension is the maximum dimensional score among all the words used in the turn. The stance strength score is the highest score of the component spurts.

For these calculations, the emotive strength scores were subdivided into three categories: **Neutral**, **Weak**, and **Strong**. The ranges for these categories are given in Figure 4.10. The black number line shows the entire [0 – 1] valence scale. This scale is subdivided into categories based on the magnitude of the distance from the neutral point. Emotive strength score ranges are given in parentheses.

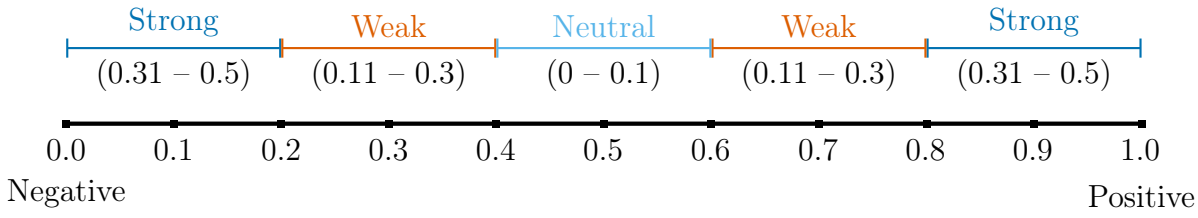
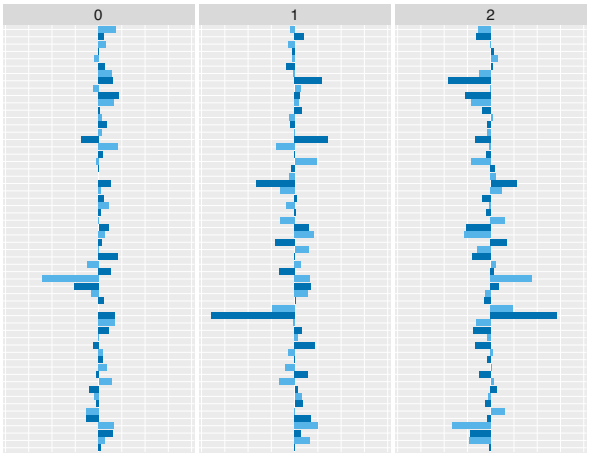


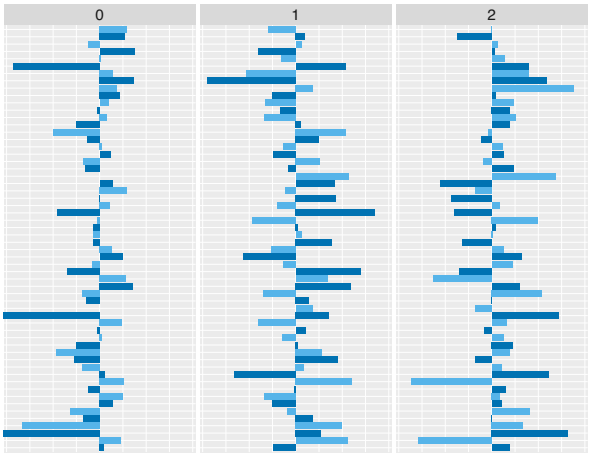
Figure 4.10: Emotive Score Category Ranges Mapped to Valence Scores

The probability of each stance strength label, conditioned on each emotive strength score range, was subtracted from the prior probability of that stance strength label. This shows the strength of the association between each emotive strength score category and stance strength. These differences are shown in Figure 4.11 for the ATAROS 6B corpus; the ATAROS 3I and PSI corpora show similar effects. Each bar represents a different speaker, however dyad pairs are adjacent. Bars extending to the left indicate that the conditional probability is greater than the prior probability, an indication that the words in that strength category are

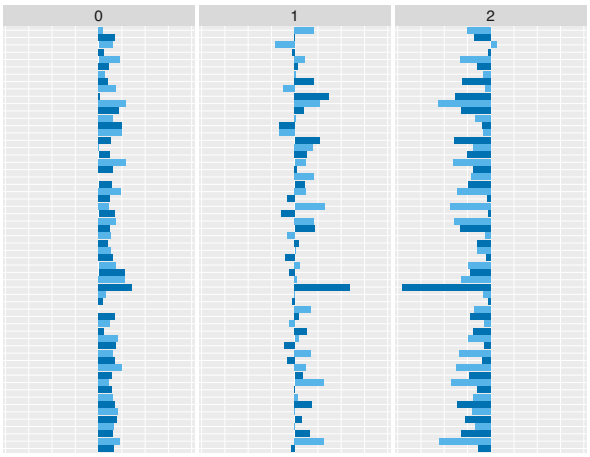
frequently used to express that specific level of stance strength. Bars extending to the right indicate that the prior probability is greater than the conditional probability, and therefore the score range is not associated with the stance strength.



(a) P(S) − P(S | WEAK)



(b) P(S) − P(S | NEUTRAL)



(c) P(S) − P(S | STRONG)

Figure 4.11: Prior Minus Conditional Probability of Stance Strength: Emotive Strength

From here the association between emotive strength and stance can be seen. While there is no consistent pattern for the **Neutral** and **Weak** score ranges, the relative magnitudes of

the bars show that the prior and conditional probabilities in the **Weak** score range are very close, a strong indication that this score range does not have much effect on stance strength. Consistent patterns emerge in the **Strong** score range. For stance strength label 0, the conditional probability is less than the prior probability universally. This pattern holds for the majority of speakers in stance strength label 1. This indicates that words with strong emotive strength are not used in stance-less, nor weak-stance spurts. For strong stance, stance strength label 2, on the other hand, the conditional probability is much higher than the prior, indicating that words with strong emotive strength are definitely used to express strong stance.

From this we can conclude that strong stance is more likely to be expressed using a strong emotive word. Strong emotive words are not used to express weak or no stance.

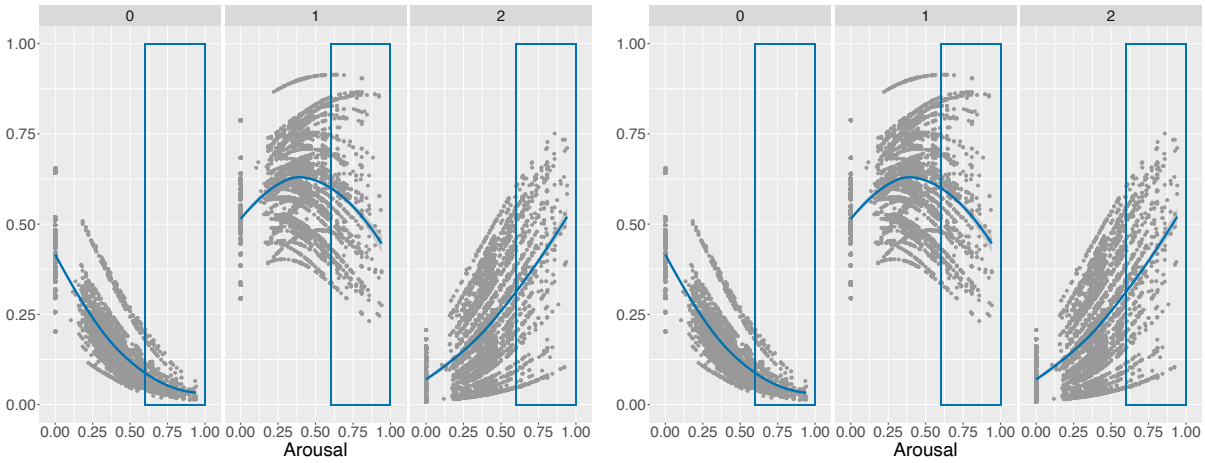
4.4 Results: Arousal

As introduced in section 4.1.2, the term *arousal* refers to the amount of excitement expressed by the word (Mohammad, 2018). Arousal is represented on a bipolar [0, 1] scale where 0 represents a complete lack of excitement (i.e. boring), and 1 represents peak excitement.

This study seeks to determine whether stance is expressed using words on the "excited" end of the arousal scale more often than on the "unexcited" end. This would be expected, particularly in a collaborative task, where much of the stance expressions would be in the context of negotiation and coming to an agreement. In trying to persuade and convince, a speaker would want to make the proposal as exciting and actionable as possible, and the use of excited language would be the most obvious way to do so.

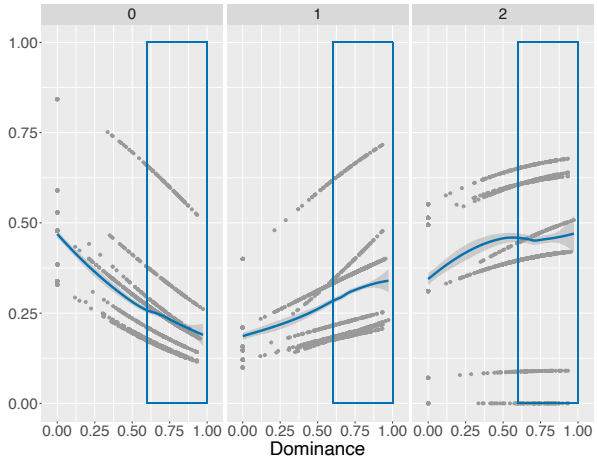
To test the effect of scores along the dimension of arousal on stance strength, multinomial logistic regression models were trained using the `multinom` function from the R (R Core Team, 2015) library `nnet` (Venables and Ripley, 2002). One model was trained for each corpus. These models use **stance** as the dependent variable, and **arousal** score and **speaker** as joint independent variables. Against a model using only speaker as the independent variable, all three corpora show statistical significance ($p < 0.001$) in a Likelihood Ratio Test

using the function `lrtest` from the R library `lmtest` (Zeileis and Hothorn, 2002). Figure 4.12 shows the fitted values generated from these models; these show how the probability of each stance strength changes as the arousal score increases. The arousal score is on the x-axis, and the probability of the stance strength label is on the y-axis. Lines were fitted using the LOWESS algorithm, the details of which are given in Section 4.3. The blue box highlights the "excited" end of the range, arousal scores of 0.6 and greater.



(a) ATAROS 3I

(b) ATAROS 6B



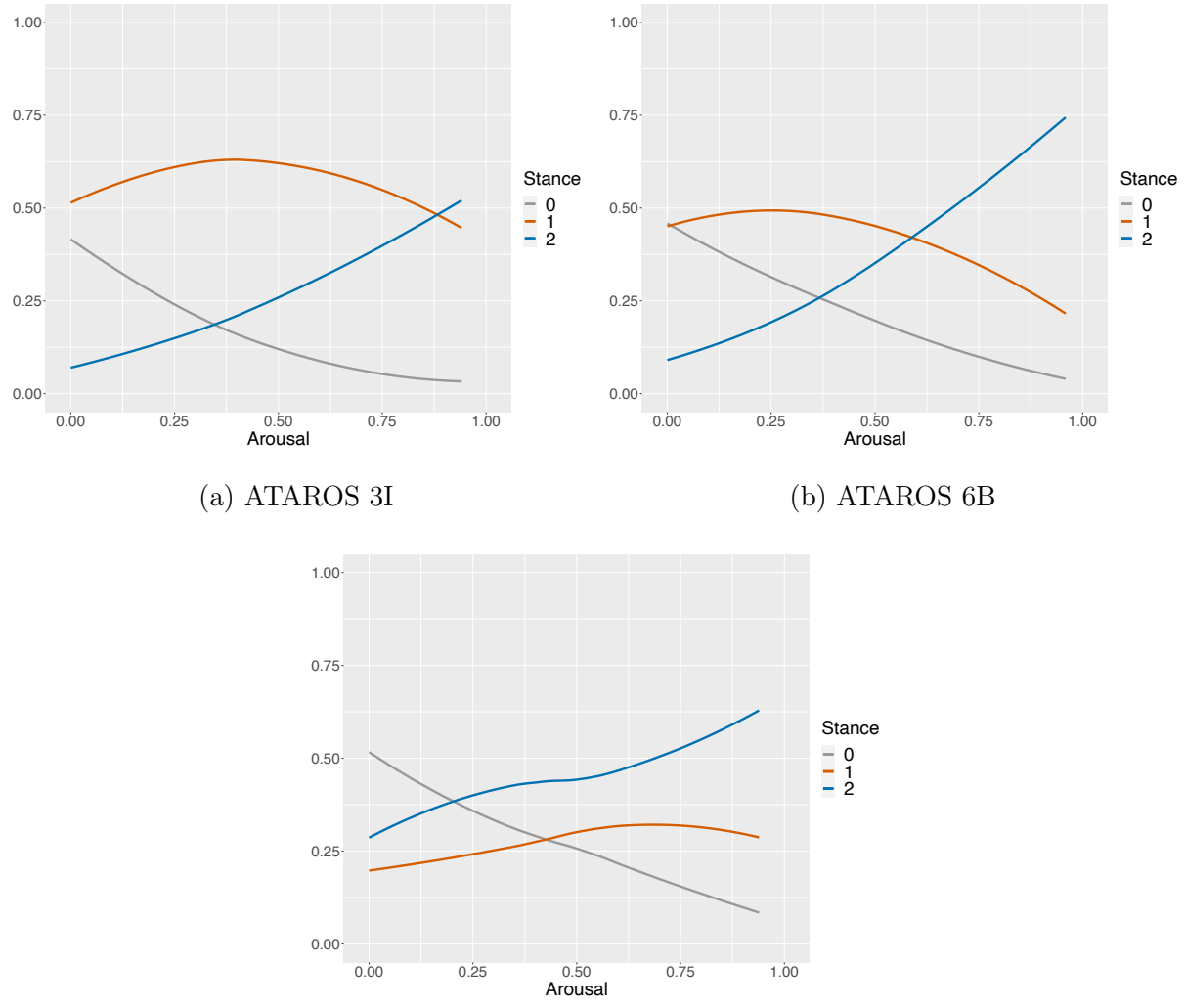
(c) PSI

Figure 4.12: Stance Strength as a Function of Arousal Score

Looking at the fitted line, it is obvious that the probability of stance strength label 0 decreases, and the probability of strong stance, stance label 2, increases, as the arousal score increases for all corpora. The ATAROS corpora show that stance strength label 1 reaches a maximum point somewhere in the "unexcited" range before decreasing. The PSI corpus, on the other hand, peaks in the excited end of the scale, around 0.7, before decreasing.

Recall that arousal is represented on a $[0-1]$ scale, with scores below 0.5 representing the "bored" or "unexcited" end of the scale, and scores above 0.5 indicating excitement. Scores of 0.5 represent neutrality, neither excited nor unexcited.

Figure 4.13 shows the fitted lines on s single plot to that we can compare the stance strengths more easily.



(a) ATAROS 3I

(b) ATAROS 6B

(c) PSI

Figure 4.13: Stance Strength as a Function of Arousal: Fitted Lines

When you compare the stance strength annotations within a single corpus, it is clear that stance strength label 1 dominates the entire ATAROS 3I corpus, except at the very peak of excitement, where 2 overtakes it. In the ATAROS 6B corpus stance label 2 begins to dominate once we reach scores in the "excited" end of the range. This applies to the PSI corpus at an even lower point; arousal scores of 0.2 and above are more likely to express strong stance than weak.

Note also that the arousal scores are concentrated at 0.2 and above; very few words at the extreme end of boredom were used. Those points at arousal score 0 are more likely due to the spurt not having any vocabulary from the NRC-VAD Lexicon.

From this, we can conclude that excited vocabulary, as represented by the arousal score, is most likely to be perceived as expressing strong stance.

To directly address the question of whether stance is more likely to be expressed using excited words, the conditional probability of each stance strength label, conditioned on arousal score, was compared to the prior probability of the stance strength label. For these calculations, the arousal scores were divided into three score ranges, **Unaroused**, **Neutral**, and **Aroused**, the ranges of which are given in Figure 4.14.

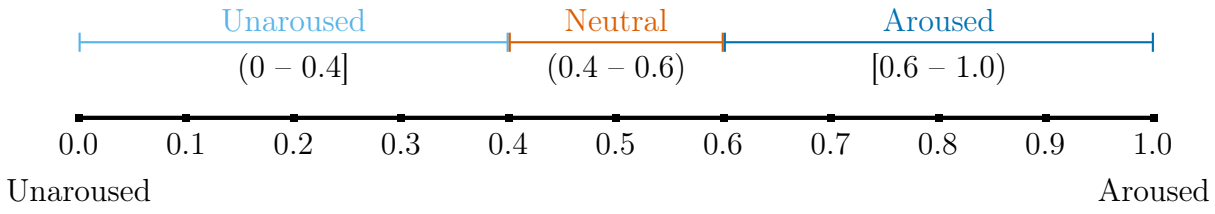
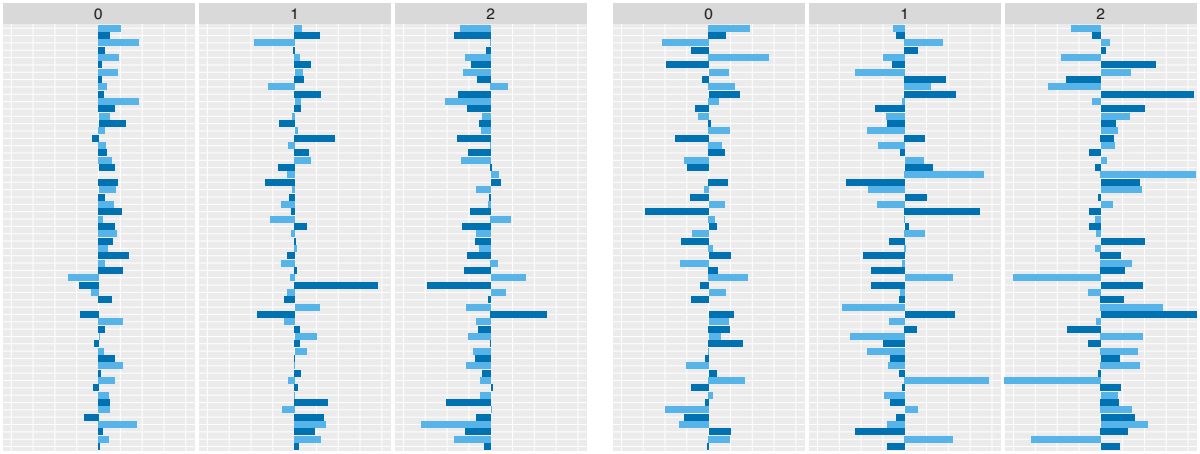


Figure 4.14: Arousal Score Category Ranges

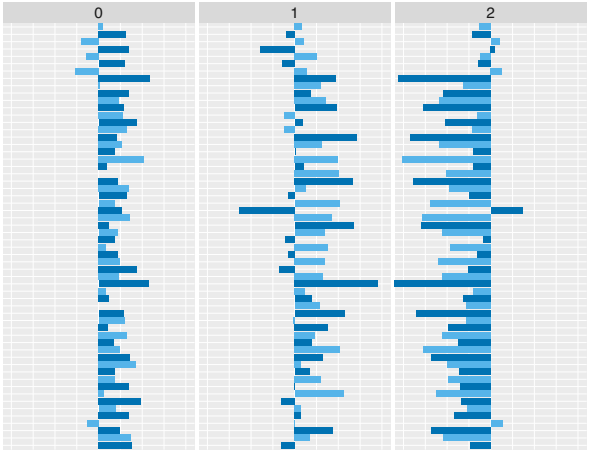
The probability of each stance strength label, conditioned on each arousal score range, was subtracted from the prior probability of the stance strength. This shows the strength of the association between stance strength and arousal score. Figure 4.15 shows the differences for the ATAROS 6B corpus; ATAROS 3I and PSI show similar effects. As explained in Section 4.3, each bar represents a different speaker; dyad partners are represented by adjacent bars.

Calculations are done at the level of the speaking turn. Bars extending to the left indicate that the conditional probability is greater than the prior probability, indicating that the score category is associated with that stance strength, while bars extending to the right indicate that the prior is greater than the conditional, indicating that it is not.



(a) P(S) − P(S | UNAROUSED)

(b) P(S) − P(S | NEUTRAL)



P(S) − P(S | AROUSED)

Figure 4.15: Prior Minus Conditional Probability of Stance Strength: Arousal

In the Unaroused category, there is no definitive pattern at stance strength 0, other than the fact that partners tend to show the same tendencies. There is a weak trend that conditional probability is greater than the prior probability at stance strength 1, indicating that, at least for most speakers, unexcited language is used in utterances expressing weak stance. The opposite is true for stance strength label 2; the prior probability is greater than the conditional probability, indicating that unexcited language is not used to express strong stance.

In the Neutral category, there is no pattern for stance strength labels 0 or 1. At stance strength label 2, for most speakers, the prior probability is greater than conditional, indicating that arousal-neutral words are not used to express strong stance.

Clearer patterns emerge in the Aroused category. Words from this category are definitely not used in stance-less nor weak stance spurts, since the prior is greater than the conditional probability for the majority of speakers, while they are definitely used to express strong stance.

From this, we can conclude that there is a strong tendency to use excited language to express strong stance. Lower levels of arousal scores are indicative or weak or no stance.

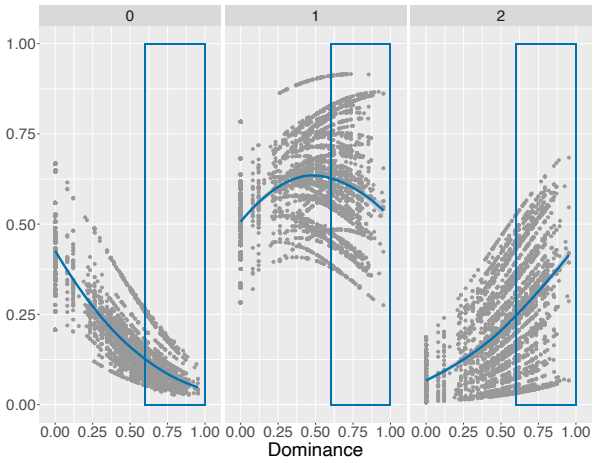
4.5 Results: Dominance

As described in section 4.1.3, the term *dominance* refers to the *dominant* $\leftrightarrow$ *submissive* or *in control* $\leftrightarrow$ *controlled* dimension of a word. It is represented on a [0,1] scale where 0 represents maximally submissive and 1 represents maximum dominance.

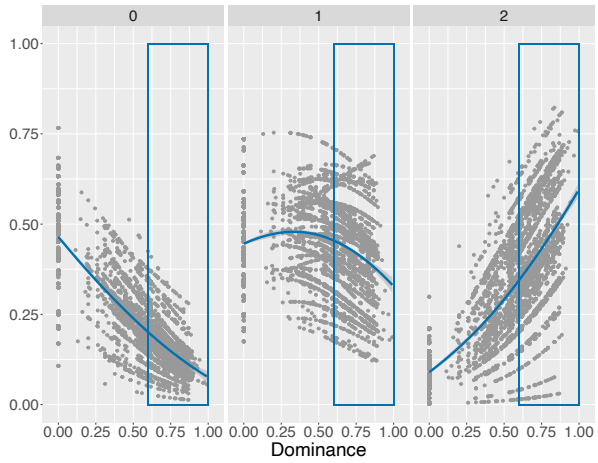
This study seeks to determine whether stance is expressed using dominant language more often than neutral or submissive language. This would be expected in a goal-oriented task. When a speaker is offering suggestions toward the completion of the task, it can be said that they are taking a leadership role in the group. These suggestions are inherently stance-laden.

To test the effect of dominance score on stance strength, multinomial logistic regression models were trained using the `multinom` function from the R (R Core Team, 2015) library `nnet` (Venables and Ripley, 2002). As described in Section 4.3, one model was trained

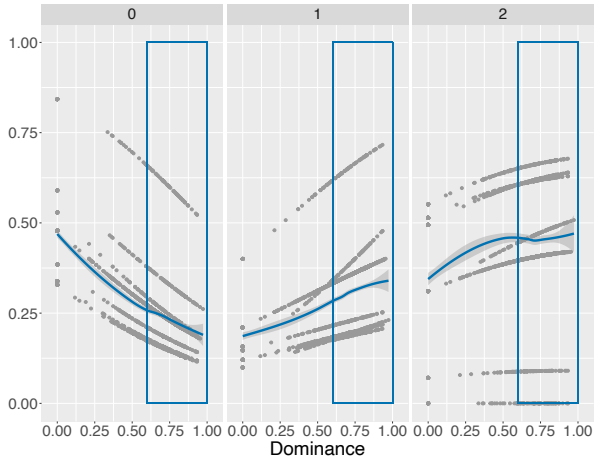
for each corpus and these models use **stance** as the dependent variable and **dominance** score and **speaker** as joint independent variables. Against a model using only speaker as the independent variable, all three corpora show statistical significance ($p < 0.001$) in a Likelihood Ratio Test using the function `lrtest` from the R library `lmtest` (Zeileis and Hothorn, 2002). Figure 4.16 shows the fitted values returned from these models. They represent the probability of each stance strength as the dominance score increases. The dominance score is on the x-axis, and the probability of the respective stance strength is on the y-axis. Lines were fitted using the LOWESS algorithm which is described in Section 4.3. The blue box highlights the dominant end of the score range, scores of 0.6 and higher.



(a) ATAROS 3I



(b) ATAROS 6B



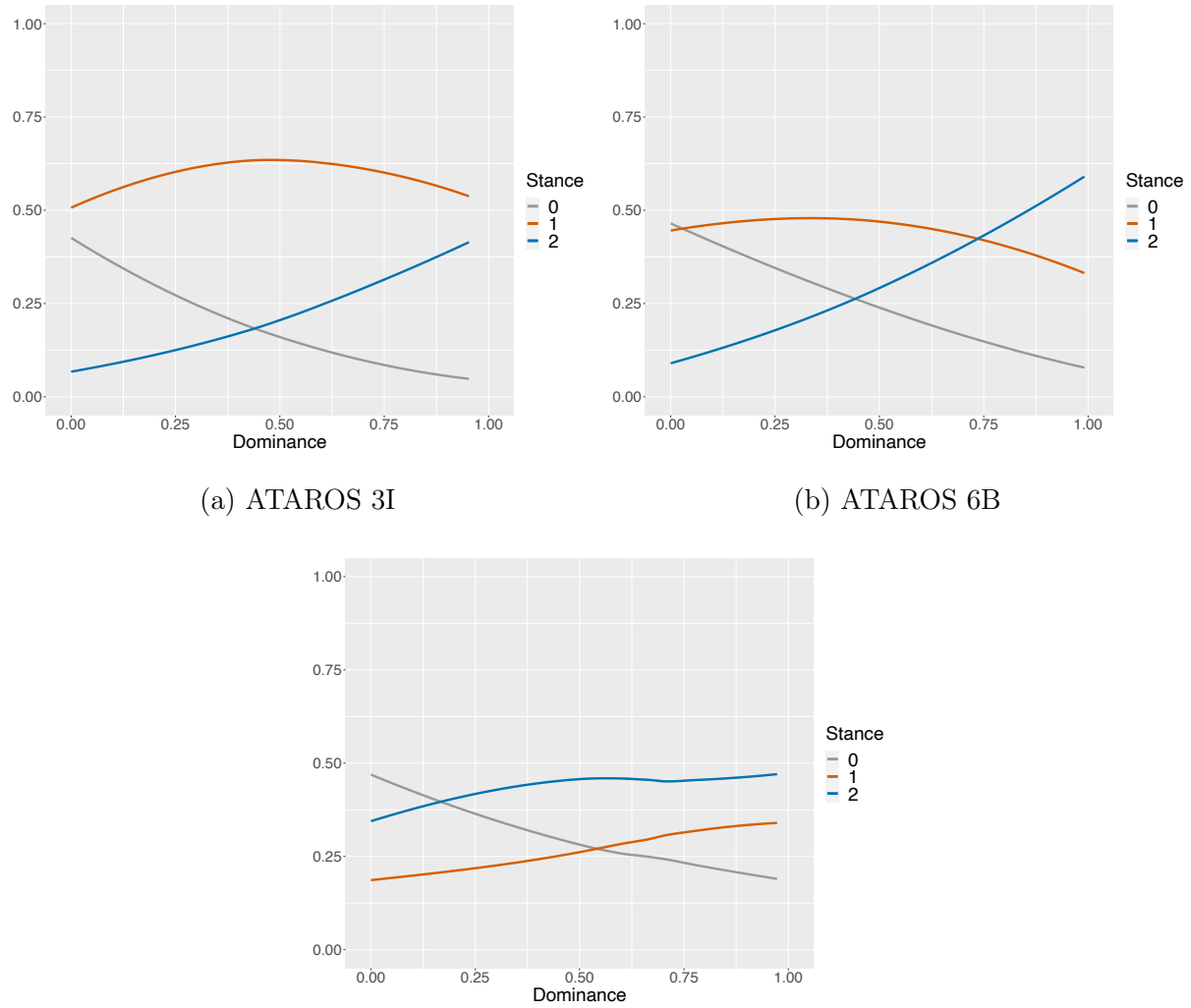
(c) PSI

Figure 4.16: Stance Strength as a Function of Dominance Score

As with emotive strength and arousal scores, the fitted line shows that the probability of stance strength 0 decreases and the probability of stance strength 2 increases as the dominance scores increase. The ATAROS corpora show that stance strength 1 reaches a maximum around the midpoint before decreasing, while the PSI corpus shows a monotonic increase throughout.

Recall that dominance scores are represented on a [0 – 1] scale, where 0.5 represents neutrality, neither dominant nor submissive. The highlighted area, therefore, represents the dominant end of the scale, scores of 0.6 and above.

Figure 4.17 shows the fitted lines on a single plot so that they can be compared directly.



(a) ATAROS 3I

(b) ATAROS 6B

(c) PSI

Figure 4.17: Stance Strength as a Function of Dominance Score: Fitted Lines

For the ATAROS 3I corpus, it is clear that weak stance dominates regardless of the dominance score. As for the ATAROS 6B corpus, stance strength label 1 dominates until scores reach the upper levels of the dominance score range. The PSI corpus shows strong stance dominating throughout the score range.

From this, we can conclude that, for high engagement tasks, dominant language increases the probability that the utterance is perceived as expressing strong stance. At lower levels of the dominance score range, weak stance is most likely.

Next, to address the issue of whether stance is more likely to be expressed using dominant language, the dominance scores were divided into three categories, **Submissive**, **Neutral**, and **Dominant**. The ranges for these categories are given in Figure 4.18.

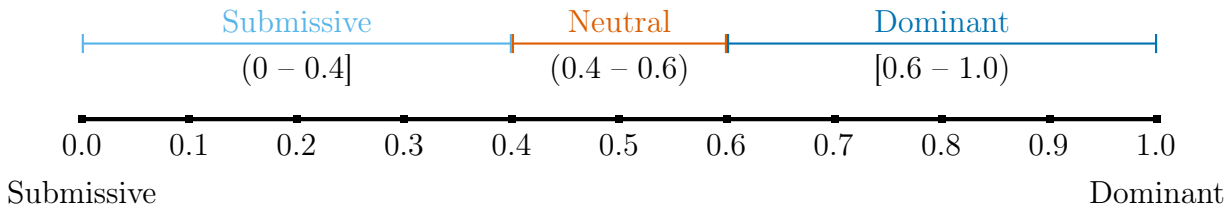
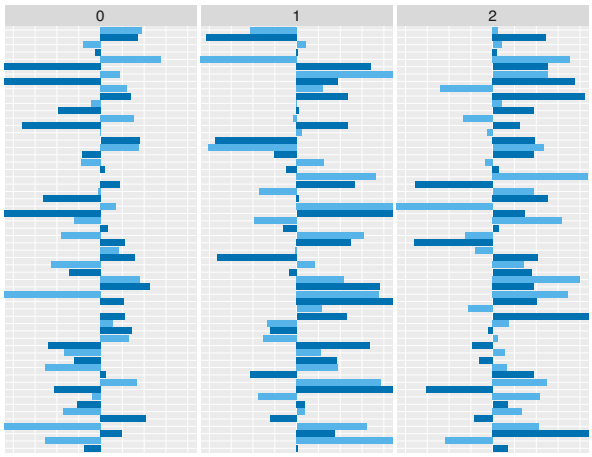
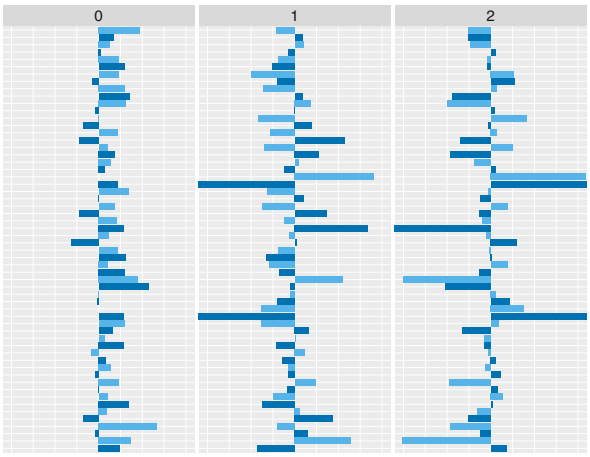


Figure 4.18: Dominance Score Category Ranges

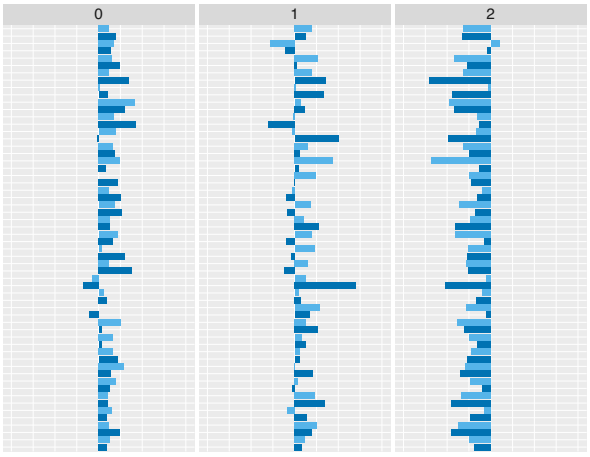
To determine whether dominant language is more likely to be used to express stance, the conditional probability of each stance strength label, conditioned on these dominance score ranges, was subtracted from the prior probability. This shows the effect each dominance score range has on the strength of the stance expressed. Figure 4.19 shows these differences for the ATAROS 6B corpus. The ATAROS 3I and PSI corpora show similar trends. Again, each bar represents a different speaker and dyad pairs are represented on adjacent bars. Bars extending to the left indicate that the conditional probability is greater than the prior probability, an indication that the score range is associated with that stance strength label, while bars extending to the right indicate the opposite, that the conditional probability is less than the prior probability, and the score range is not associated with the stance strength.



(a) P(S) − P(S | SUBMISSIVE)



(b) P(S) − P(S | NEUTRAL)



(c) P(S) − P(S | DOMINANT)

Figure 4.19: Prior Minus Conditional Probability of Stance Strength: Dominance

In the Submissive category, there are tendencies showing that submissive language is not strongly associated with stance, since the prior probability is greater than the conditional for stance strength 1 and 2, but this is far from universal. The magnitude of the bars shows that the conditional probability is very low. As for the Neutral category, the prior probability is greater than the conditional for stance strength 0, indicating that dominance-

neutral language is not used in stance-less statements. There is no discernible pattern for stance strenghs 1 and 2.

Universal patterns emerge in the Dominant category. Stance-less and weak stance utterances do not include strongly dominant words, while expressions of strong stance do.

From this, we can conclude that dominant language is used to express strong stance.

4.6 Dimensional Alignment

Throughout this investigation, I was picturing a scenario in which one speaker's dimensional strength increased the level of dimensional strength by their speaking partner. It stands to reason that strong emotion would be met with strong emotion, either along the same polarity in agreement, or along the opposite polarity in disagreement. Similarly, it would not be outside of the realm of possibility that excitement would yield excitement, or dominance would yield dominance if both personalities were trying to lead the discussion. As a second research question, then, I decided to test this theory. I am using the term alignment as opposed to entrainment; entrainment would indicate a near match in scores, while alignment means that both scores fall within the same general score-range.

Preliminary studies showed that, when a speaker chooses a word, it is more often on the basis of a single dimension, rather than a combination of all three. This is not shocking since these dimensions work independently of each other, particularly *valence* and *arousal*, as can be seen in Figure 4.5. When choosing the word that most closely matches the message a speaker wishes to convey, they will use one that most aligns with the dimension they wish to emphasize. For instance, a speaker wishing to express the positive attribute of *kindness* (valence score 0.938, emotive strength 0.438) would chose that word in spite of its low arousal score (0.350) and neutral dominance score (0.508).

To determine whether speakers align in terms of dimensional strength, particularly along the high ends of the scales (positive, excited, and dominant), the first task is to establish a measure of dimensional strength for a speaking turn. To do this, rather than the raw score, the magnitude of distance of the score from the neutral point (|0.5 - score|) is used, so that

entrainment at the weak end of the scale (negative, unexcited, submissive) is also considered. Note that this is the basis for the emotive strength score already used in Section 4.3.

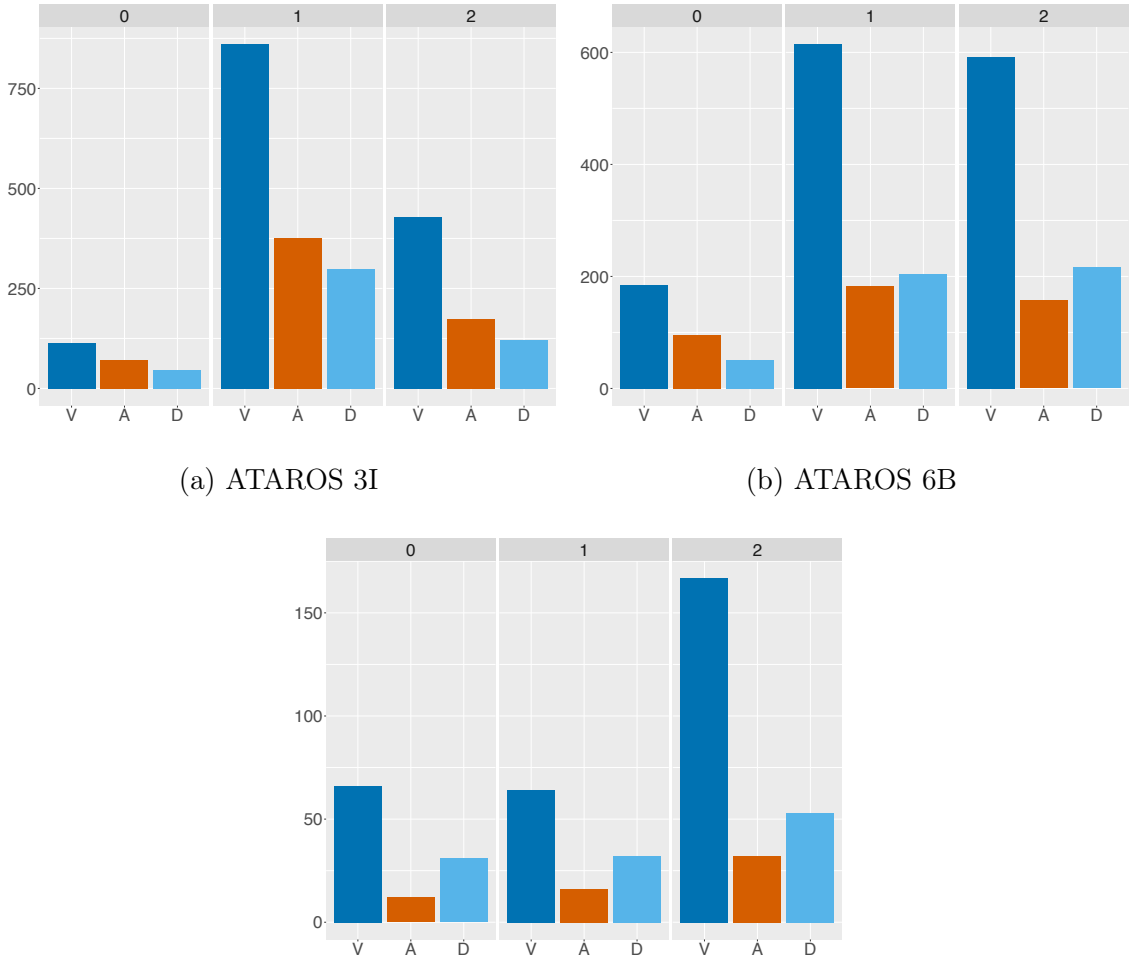
For each word from the NRC-VAD Lexicon found in a speaking turn, the strength scores along all three dimensions are compared. The dimension with the highest strength score is considered the dimensional emphasis for the turn.

As an example, in the following exchange, the word *think* has the greatest magnitude along all dimensions in speaker A's turn. This is the score along the *valence* dimension. Therefore, valence is the dimensional emphasis for speaker A's turn. Speaker B's resonse has *arousal* as its dimensional emphasis.

A: Do$\{^{V:0.670}_{A:0.548}_{D:0.536}\}$ you think$\{^{\mathbf{V: 0.786}}_{A:0.408}_{D:0.618}\}$ we'll have$\{^{V:0.757}_{A:0.389}_{D:0.593}\}$ shoes? We have shoe laces.

B: Yeah. Shoot$\{^{V:0.092}_{\mathbf{A: 0.926}}_{D:0.625}\}$.

Figure 4.20 shows that *valence* is the emphasized dimension for the majority of speaking turns at all three levels of stance across all three corpora. The ATAROS 3I corpus shows that the next most frequent dimension is *arousal*, then *dominance*. This pattern continues in the ATAROS 6B corpus for stance-less speaking turns. For stance-laden speaking turns in the ATAROS 6B corpus, and throughout the PSI corpus, *dominance* is the second most frequent dimension, followed by *arousal*.



(a) ATAROS 3I

(b) ATAROS 6B

(c) PSI

Figure 4.20: Maximum Scoring Dimension

In the context of determining whether word choice is influenced by the words used by our speaking partner in terms of emotional dimensions, it makes sense to restrict our study to blocks of text in which both speakers are using words from the NRC-VAD Lexicon in sequential speaking turns. Within these blocks of dialogue, for each turn, the dimensional scores of the current turn are compared to those of the previous since these are the words that are most likely to influence a speaker's lexical choices.

Going back to the previous example, while *arousal* formed the dimensional emphasis for speaker B's turn, for the purpose of measuring dimensional alignment, the maximum *valence* score from speaker B's turn is going to be compared with the *valence* score of speaker A's turn. *Arousal* will then be the dimension forming the basis of the comparison to the next speaking turn. Had there been more than one emotive word in speaker B's turn, the highest *valence* score among them would have been used.

Once the emphasized dimension of the previous speaking turn has been identified, the original [0 – 1] score for that dimension in the current turn is compared to the score in the previous turn. Figure 4.21 shows the scatterplot for a combined ATAROS 3I and 6B corpus; as individual corpora, both looked very similar. Figure 4.22 shows the PSI corpus. These represent turn-by-turn dimensional score matches. The centroid of the data for each dimension is represented on the scatterplot with an asterisk (*). The plot is divided into four quadrants, as in a Cartesian plane, with a horizontal and a vertical line at the neutral point, 0.5, along both axes. A point in the (x, y) quadrant would represent a case where positive/excited/dominant language formed the dimensional emphasis for the previous utterance and positive/excited/dominant language was used in response. A point in the (x, -y) quadrant would represent where positive/excited/dominant language was met with negative/unexcited/submissive language. Density in the (x, y) and the (-x, -y) quadrants represent dimensional alignment at the high and low ends of the scales. I say alignment rather than entrainment because evidence of entrainment would have shown density concentrated along a diagonal, representing y = x, which is not the case here.

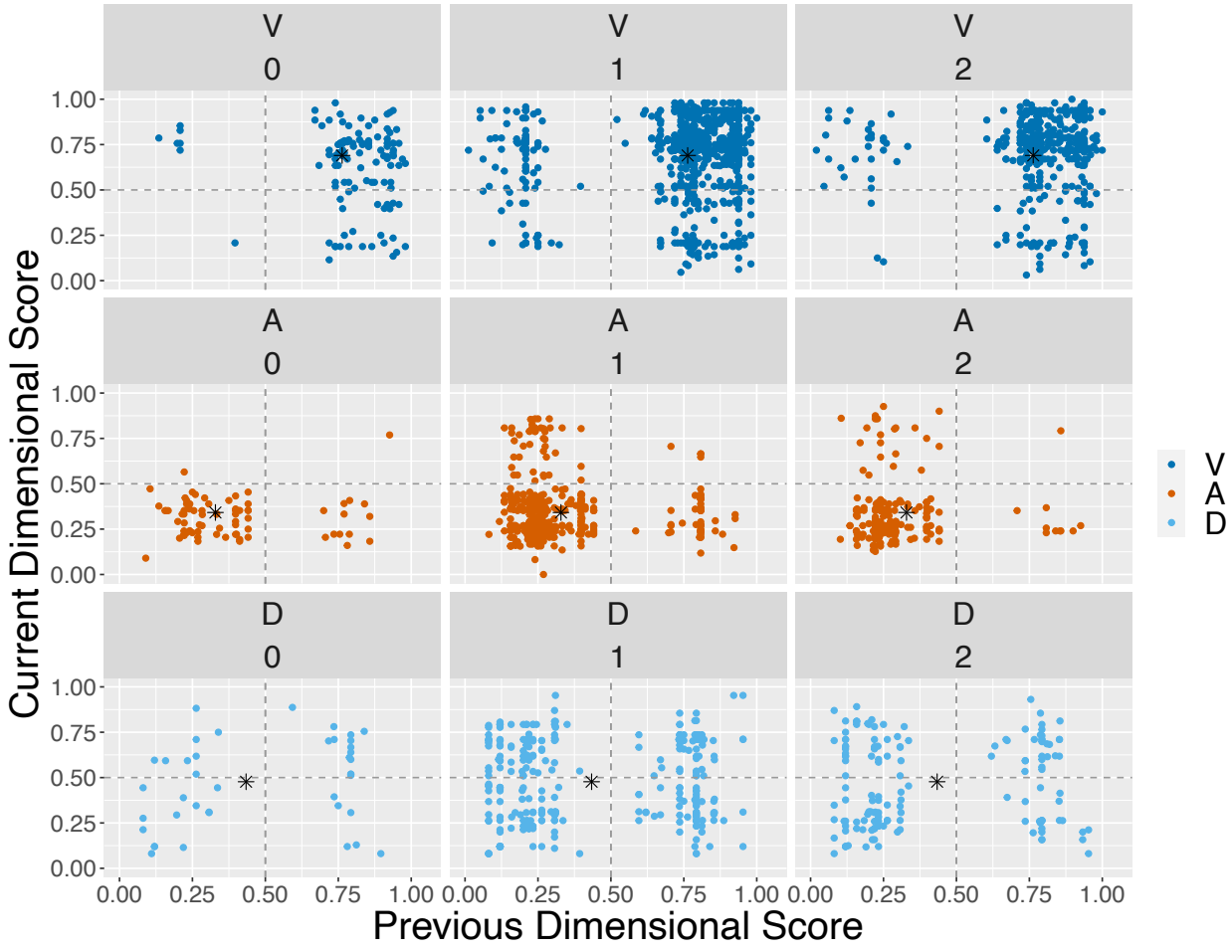


Figure 4.21: Dimensional Alignment: ATAROS 3I & 6B Corpora

For the ATAROS corpora, points are concentrated in the (x, y) quadrant for all three levels of stance for the *valence* dimension, showing alignment at the positive end. As for *arousal*, density is concentrated in the (-x, -y) quadrant, again at all levels of stance, indicating alignment at the "unexcited" end of that scale. The lack of concentration of points for *dominance*, and the fact that centroid is near the origin, speaks to the fact that there is no alignment along this dimension.

The y-values of the centroids show the effect of stance strength on this dimensional alignment. There is little difference between the different levels of stance strength.

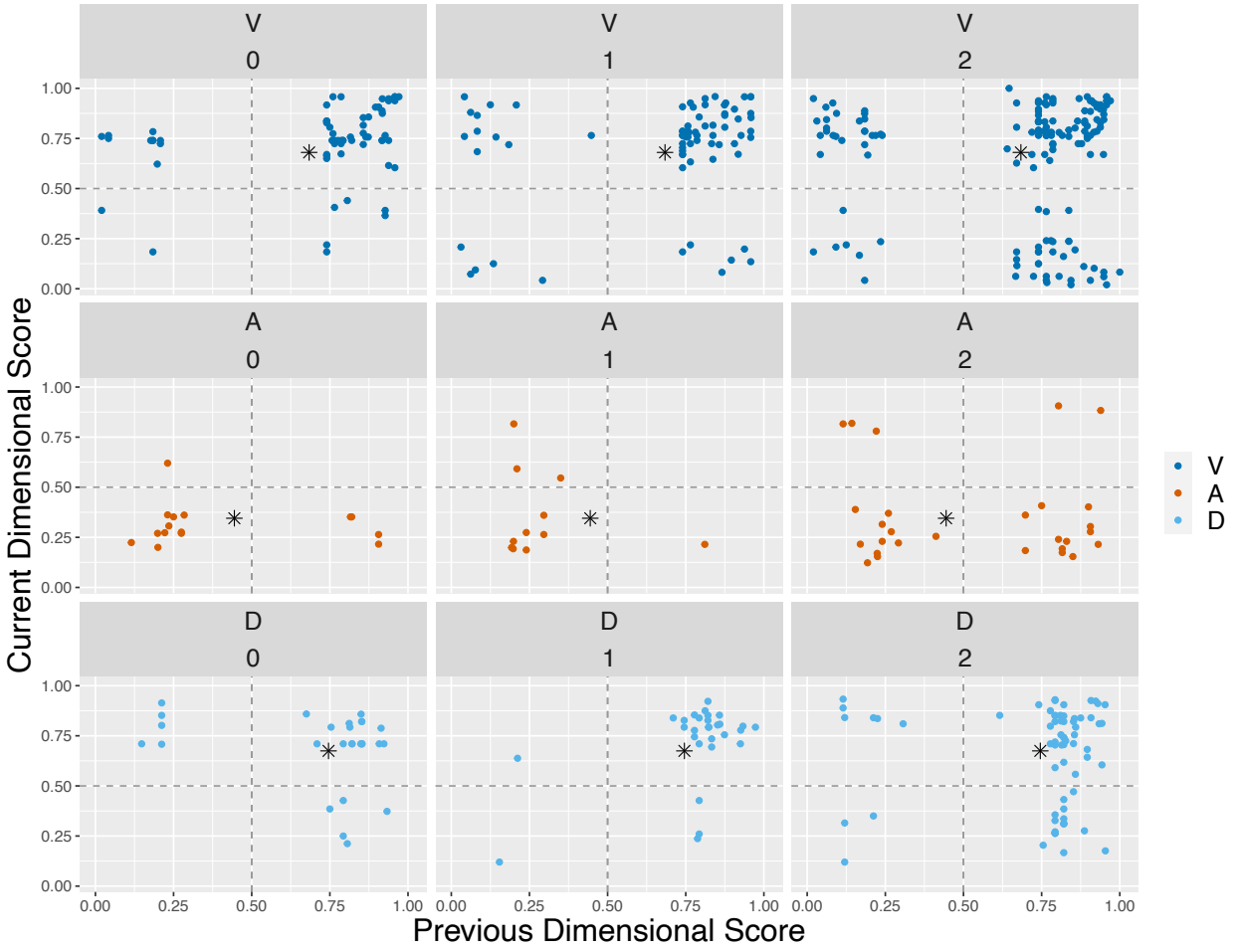


Figure 4.22: Dimensional Alignment: PSI

As for the PSI corpus, again it shows alignment at the positive end of *valence*. Data sparsity for *arousal* hinders drawing any conclusions for this dimension. There is, however, indication of alignment at the dominant end of the *dominance* dimension.

Again, the y-values of the centroids show no effect of stance strength on dimensional alignment.

From this, I conclude that, while there is alignment along the positive end of the valence dimension, and the "unexcited" end of arousal, this cannot be considered entrainment. Had it been entrainment, the dimensional scores would have matched in magnitude, and the points would have been concentrated more along the diagonal.

The alignment in *valence* is expected due to our tendency toward pleasantness and lack of confrontation in discourse; even in disagreements, there is an attempt to frame differences in positive terms, and to mitigate the effect of face threatening acts (Lakoff, 1973). Since there was no difference in rates of alignment between different levels of stance strength, stance has no effect on this alignment. As for *arousal*, the laboratory setting, the fact that dyad pairs had only just met, and the artificiality of the tasks may have lessened any alignment effects in the ATAROS tasks. The tendency was toward alignment in the "unexcited" end, though one might expect in discourse between close friends, to see alignment at the "excited" end.

Within the context of a senatorial hearing, there was little to no alignment on the *arousal* dimension, however, there was at the dominant end of the *dominance* dimension. This should not surprise given the parties involved and the inherent adversarial nature of the hearing. Again, however, there appears to be no effect of stance strength.

4.7 Discussion

This study seeks to determine whether there is a relationship between emotional strength, along three dimensions, and the perceived strength of the stance expressed. Additionally, since the focus is on the dialogical nature of stance taking, there was an investigation into alignment along these discrete emotional dimensions.

For all three corpora, I have shown that there is a relationship between the magnitude of *valence* score and the strength of the stance expressed. Strong stance is expressed using strongly negative or strongly positive words. Lower levels of emotional strength are used to express weak or no stance. A good example of the use of strong words is in the following exchange:

A: Betcha we could cut public news station and keep$_{\{V:0.714\}}$ the public access - access station (1).

B: So that - boy, I really hate$_{\{V:0.031\}}$ cutting that - the - (2)

Compare speaker A's utterance anchored by the word *keep*, whose valence score of 0.714 yields a magnitude of 0.214. It was annotated with stance strength label 1. Contrast that with speaker B's utterance, anchored by the word *hate*, valence 0.031, magnitude 0.469, which was annotated with stance strength 2 likely on the strength of the word *hate*. As a matter of fact, across all three corpora, the word *hate* appears in 13 spurts. Only two of the thirteen spurts were annotated with stance strength label 1, and none with 0. Both instances of weak stance annotation were in an aside such as "I hate to do this but.."

A conditional probability calculation further enforces this connection. When the prior probability of each stance strength label is compared to the probability conditioned on varying levels of valence scores, clear patterns emerge. Strongly emotive words are consistently used to express strong stance and rarely to express weak or no stance. At lower levels of valence strength, there are no clear patterns.

Similar patterns emerge with *arousal* scores, which is noteworthy since Figure 4.5 shows that these dimensions are largely independent of each other. In tasks involving high levels of engagement, high arousal scores are indicative of strong stance. Notice how the use of the word *chaos* adds to the strength of the stance expressed by speaker B.

A: Taxi stops? We probably don't need$_{\{A:0.574\}}$ to do that either. (1)

B: Well, otherwise, it could be chaos$_{\{A:0.923\}}$, with taxis stopping everywhere. If you have stops, it could help a little bit. (2)

Interestingly, as strong as the word *chaos* is along the *arousal* dimension (emotive strength score: 0.423), its *valence* score of 0.016 (emotive strength score: 0.484) is even stronger.

As for scores along the *dominance* dimensions, these are also shown to have an effect on the strength of the stance expressed. This is not surprising given the alignment shown between *dominance* and *arousal* in Figure 4.7. These studies have shown that, for high engagement tasks, dominant language is indicative of strong stance, and for all tasks, dominance scores of 0.5 and above are indicative of stance.

The next inquiry was to rule out the scenario where an increase along these dimensions was due to an increase in emotions. This, in essence, is a description of dimensional entrainment. I was not able to find evidence of dimensional entrainment, but I did find dimensional alignment. Speakers were found to align along the positive end of the *valence* dimension, and the "unexcited" end of the *arousal* dimension in non-formal speech. This is consistent with theories of linguistic entrainment, and sociolinguistic theories such as Bell's Audience Design, and Communication Accommodation Theory (CAT) (Giles and Powesland, 2009), both of which address ways in which speakers adjust their speaking style, or behaviour, to show solidarity and build rapport with their speaking partner. This is particularly important in undertaking a collaborative task. Interestingly, given that these tasks were focused on eliciting stance, and involved negotiation, there was no entrainment along the *dominance* dimension. The inter-speaker rapport was dominantly co-operative.

In a more formal setting, the PSI hearings showed alignment along the dimensions of *valence* and *dominance*, though there was less need for rapport building. This was likely more due to the formality of the setting and the standards of behaviour expected, and the parties involved, to have reached the positions they held (United States Senators and a Wall Street executive), have adopted dominance-heavy speaking tendencies. This is evident from the score distribution given in Figure 4.18.

This chapter shows that word strength along the dimensions of *valence*, *arousal*, and *dominance* are correlated with the perceived strength of the stance being expressed. While it shows a tendency toward alignment along these dimensions, there was no evidence that one speaker's use of a strong word caused their speaking partner to match strength along the same dimension. Instead, the tendency was for speakers to show strength along the dimension of their choice, independent of their speaking partner.

While the strength of the lexicon we use is something speakers are somewhat conscious of, the following chapters address dimensions of dialogical cohesion in aspects speakers may be less aware of. In Chapter 5, I address how speaking partners converge in the terminology they use to refer to ideas and entities within a discourse. In Chapter 6, I discuss what is argued

to be a largely unconscious dimension (Danescu-Niculescu-Mizil et al., 2012), convergence in function word use.

Chapter 5
TERMINOLOGICAL ALIGNMENT

In the course of a dialogue speakers must establish between them a set of mutual beliefs, assumptions, and knowledge about the content of the dialogue, the environment in which it is happening, and the participants, both those who are actively speaking and those who may be just observing. This is known as **common ground**. This is a collaborative effort (Clark and Wilkes-Gibbs, 1986; Garrod and Anderson, 1987; Brennan and Clark, 1996) and is often reflected in the level of engagement shown by the participants. When a speaker introduces a new topic, they also propose a conceptualization of that topic through the words they choose. If the speaking partner adopts this terminology, they are signaling that they also accept this conceptualization (Garrod and Anderson, 1987). Where the term is not acceptable or understood, the speaking partner can seek confirmation or clarification, or reject the term. In this rejection, a new, more suitable term may be proposed. Once a term has been accepted by both parties, the speakers enter into a *conceptual pact* (Brennan and Clark, 1996).

Common ground and shared conceptualizations are often signaled through **dialogic resonance**, which Du Bois (2014) defines as "catalytic affinities across utterances." This refers to a sense of connectedness between utterances through the use of shared or similar vocabulary, syntactic structure, prosody, or any other level of linguistic abstraction. Through these resonances, a speaker is signaling that they understand and acknowledge what others have said (Nir and Zima, 2017).

The most straightforward form of dialogic resonance involves a repetition of parts, or even a paraphrase, of the previous speaker's utterance (Sakita, 2006). Incorporating the content and proposed conceptualization into their own utterance serves as an even stronger

signal. This goes beyond simple repetition and parroting. The use of a parallel structure where there are other options available holds meaning outside of the literal meaning of the words (Du Bois, 2014).

As an example, the following excerpt is from the ATAROS 3I task. The participants are trying to decide where to place an item that does not fit tidily into their organizational scheme.

A: We can put it where we want, though. We're in charge here. So - (2)

B: We can put it wherever we want. (1)

Note the parallelism between these utterances, made more clear in the following alignment, known as a digraph (Du Bois, 2014):

A:	We can put it	where	we want	though	.
B:	We can put it	wherever	we want		.

Speaker B could have responded to speaker A's utterance in many other ways, such as a simple agreement ("Yes" or "Yes we can"), or they could have chosen not to respond to it directly at all. Instead, they opted to to express agreement and solidarity with their speaking partner through the use of a parallel structure; they are indicating that they are in agreement regarding the nature of the ATAROS task.

Parallel structures are an effective way to express dissent or opposition (Sakita, 2006; Du Bois, 2014). The following is from a rather contentious exchange between Lloyd Blankfein and Carl Levin during the US Senate Homeland Security Permanent Subcommittee on Investigations (PSI) hearing on the 2007 – 2008 financial crisis. Lloyd Blankfein is trying to justify the decisions made by Goldman Sachs, which Carl Levin is arguing put them in a conflict of interest situation:

LLOYD BLANKFEIN: You know, we live in different contexts and this is a professional - (1) this is a - a mark- (1)

CARL LEVIN: Maybe just call it a human context. (2)

Here, Lloyd Blankfein is referring to the professional, detached context in which his company made their decisions leading up to the financial crisis. Carl Levin is contrasting this with *human* context to argue that these decisions were unfair to customers. The parallelism between these structure is made more clear in the following digraph:

Lloyd Blankfein:	.. we live in	different	contexts	.. (1)
Carl Levin:	.. let's call it	a human	context	. (2)

A more syntactically opaque instance of resonance appears in the following.

CARL LEVIN: By the way they have an idea, more than an idea in these cases. But putting that aside, what do you think about (1) selling securities which your own people think are (1) crap? .. Does that bother you? (2)

LLOYD BLANKFEIN: I think they would - (1) again, as a hypothetical? (0)

CARL LEVIN: No, this is real. (2)

Syntactically, these utterances are very different, but the use of the contrasting terms *hypothetical* and *real* makes the resonance between them clear.

Lloyd Blankfein:	.. again, as a	hypothetical	? (0)	
Carl Levin:	No, this is	real	. (2)	

This example highlights that not all parallel structures use the same vocabulary. For this reason, a turn-by-turn comparison of the words in sequential speaking turns would not be

sufficient to identify resonant pairs of utterances, nor would this method yield a quantifiable measure of engagement. A more abstract similarity measurement is required.

The goal of these experiments is to determine whether the level of engagement, as demonstrated by the use of parallel linguistic structures or shared conceptualizations, varies with the strength of stance being expressed. I predict that the level of engagement will be greater with stronger stance. To be able to evaluate, appraise, predict, or opine about something, particularly when done strongly or confidently, one must be focused on the topic at hand.

Before establishing this relationship, however, the first task is to establish a measurement of engagement. For this I am using Word Mover's Distance (WMD) (Kusner et al., 2015). WMD is defined as "the minimum amount of distance that the embedded words of one document need to 'travel' to reach the embedded words of another document" (Kusner et al., 2015). This calculates Euclidean distance, and treats both documents as a bag of words, with function words removed. This has the effect of giving pairs of words with similar meanings, whether they are synonyms or antonyms, a low score. Where there is not a perfect correspondence between word pairs, scores are distributed among the closest matches. Note that it is a distributional similarity measurement, and not strictly a semantic one; what is parallel in structure is likely to be parallel in meaning (Du Bois, 2014). Where a word appears in both documents, the cost to move between those words is 0.

Although this algorithm returns relatively low scores for pairs of utterances expressing semantically similar ideas, or using lexically similar phrases; it could, however, easily miss parallel structures that lack these similarities, such as those using metaphor.

These measurements will be validated against manual judgements of engagement assigned to a subset of the ATAROS corpus. I predict that lower WMD scores correspond to a higher level of engagement.

Following that, I will establish a relationship between engagement, as represented by WMD score, and stance strength. I predict that expressions of stronger stance will show lower WMD scores, indicative of a higher level of engagement, than expressions of weaker stance.

5.1 Experimental Design

Using the `gensim` (Řehůřek and Sojka, 2011) implementation of Word Mover's Distance (WMD) (Kusner et al., 2015), and the pre-trained `Google News 300 Word2Vec` model, a pairwise similarity measure was calculated for each combination of spurts from sequential speaking turns. These experiments were run on the spurt rather than the speaking turn level since a spurt corresponds roughly to an intonational phrase (Shriberg et al., 2001), which correspond to "coherent units of thought which are fully active in the mind" while they are being spoken (Chafe, 1994; Sakita, 2006).

To demonstrate the pairwise comparison of spurts, the following excerpt shows a multi-spurt speaking turn by speaker A. The spurts are numbered here only for the purpose of this example:

1 - A: Um, well, in a - in a - I'm thinking, in a real grocery store, wouldn't the shoelaces be in, like, this kind of an area? (2)

2 - B: Yeah it w- yeah, actually it would, because that's where you find shoe polish, also. (2)

3 - A: Or not? (1)

4 - A: Sorta - (x)

5 - A: Sort of shoe polish, tweezers - (1)

6 - B: Okay. (1)

Table 5.1 shows the spurts from this example that will be paired for comparison. The WMD score will be calculated and assigned to the second member of each pair. Note that the sequential spurts by speaker A, numbers 3 – 5, will each be assigned a WMD score based on a comparison with spurt number 2. Spurt number 6, on the other hand, will be assigned three separate WMD scores, one in comparison with spurt 3, one for 4, and one for 5.

	Spurt Number				
Speaker A	1		3	4	5
Speaker B		2			6

Table 5.1: Pairwise Comparison of Spurts Example

Spurts were case normalized and tokenized. Stop words, from the list included in the NLTK English corpus (Loper and Bird, 2004), were removed prior to calculating the WMD score. Spurt-level WMD scores, stance strength, and the text of the spurts were retained for further analysis.

Table 5.2 shows the count of spurts used for these experiments. Note that the PSI corpus had considerably longer speaking turns, hence the higher spurt count.

Corpus	# **Spurts**
ATAROS 3I	10435
ATAROS 6B	9504
PSI	14534

Table 5.2: Spurt Count

Figure 5.1 shows the distribution of WMD scores in the three corpora. Note the normal distributions centered around the mid-point of the [0, 5.6] range with the exception of a peak at 0 for both ATAROS tasks.

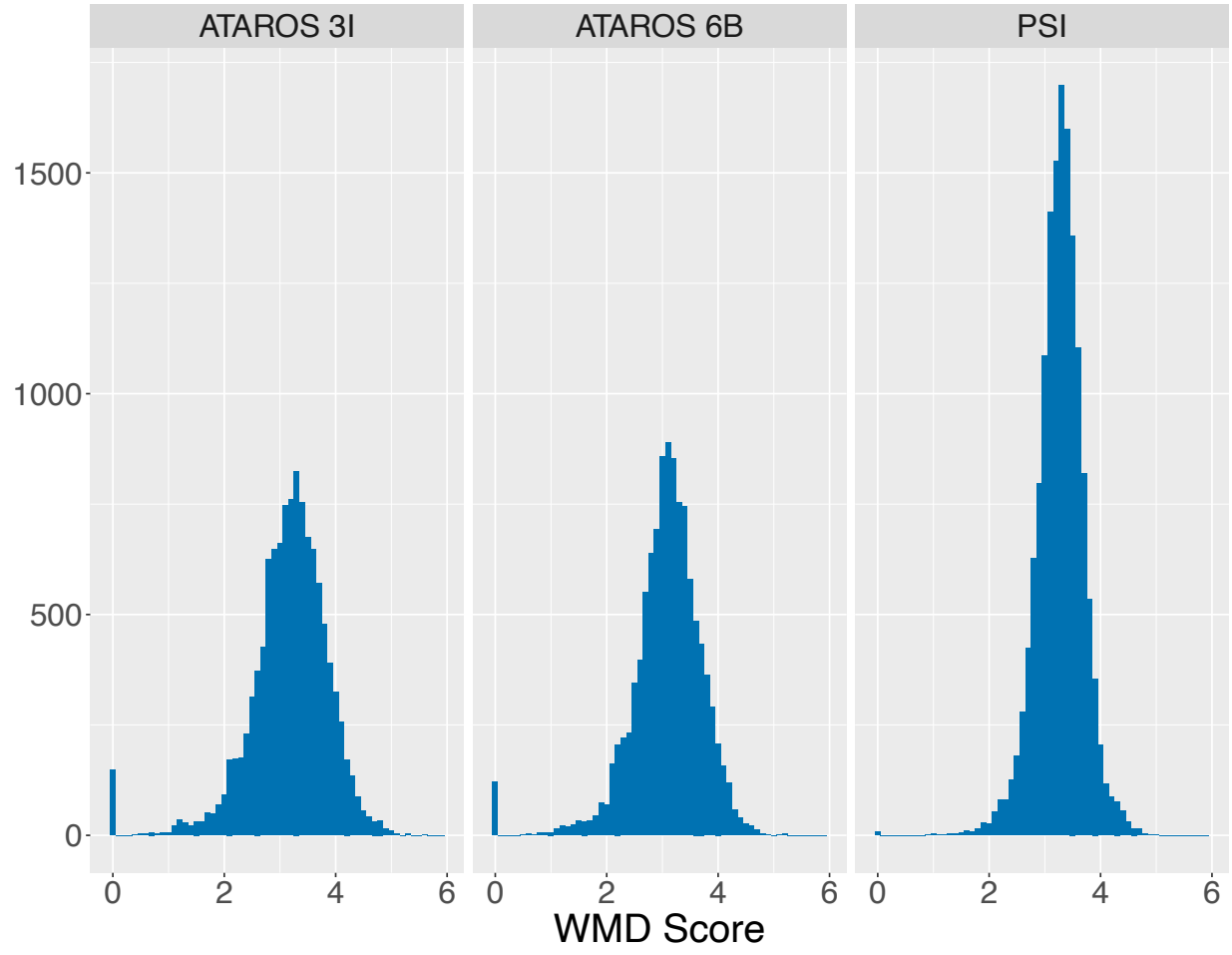


Figure 5.1: Distribution of WMD Scores

5.2 WMD Scores as a Measure of Engagement

The first set of experiments is intended to establish whether WMD scores are a good measurement of engagement.

As part of the ATAROS project, a subset of the ATAROS corpus (14 out of 30 dyads) was annotated for speaker-level engagement in both the 3I and 6B tasks. Engagement judgements were based on the general sense of how involved each speaker was in discussing the topic. The engagement annotation was assigned to the entire task; one label was assigned to each speaker for the 3I task, and one was assigned for the 6B task. One speaker per task was

annotated with low engagement.

Table 5.3 shows the breakdown of spurts annotated for engagement.

	High	**Low**
ATAROS 3I	5048	216
ATAROS 6B	4235	200

Table 5.3: Spurt Level Engagement Annotations

A Pearson χ^2 Test (Pearson, 1900) shows a cross-task statistically significant relationship ($p < 0.005$) between WMD scores and human-annotated engagement scores. On the corpus level, however, only the 6B corpus shows significance ($p < 0.005$). For this reason, this experiment will be run on a combined ATAROS 3I and 6B corpus; only spurts from those dyads annotated for engagement will be considered.

Figure 5.2 shows how the WMD scores vary between the two levels of engagement. An **x** on each box indicates the mean WMD score.

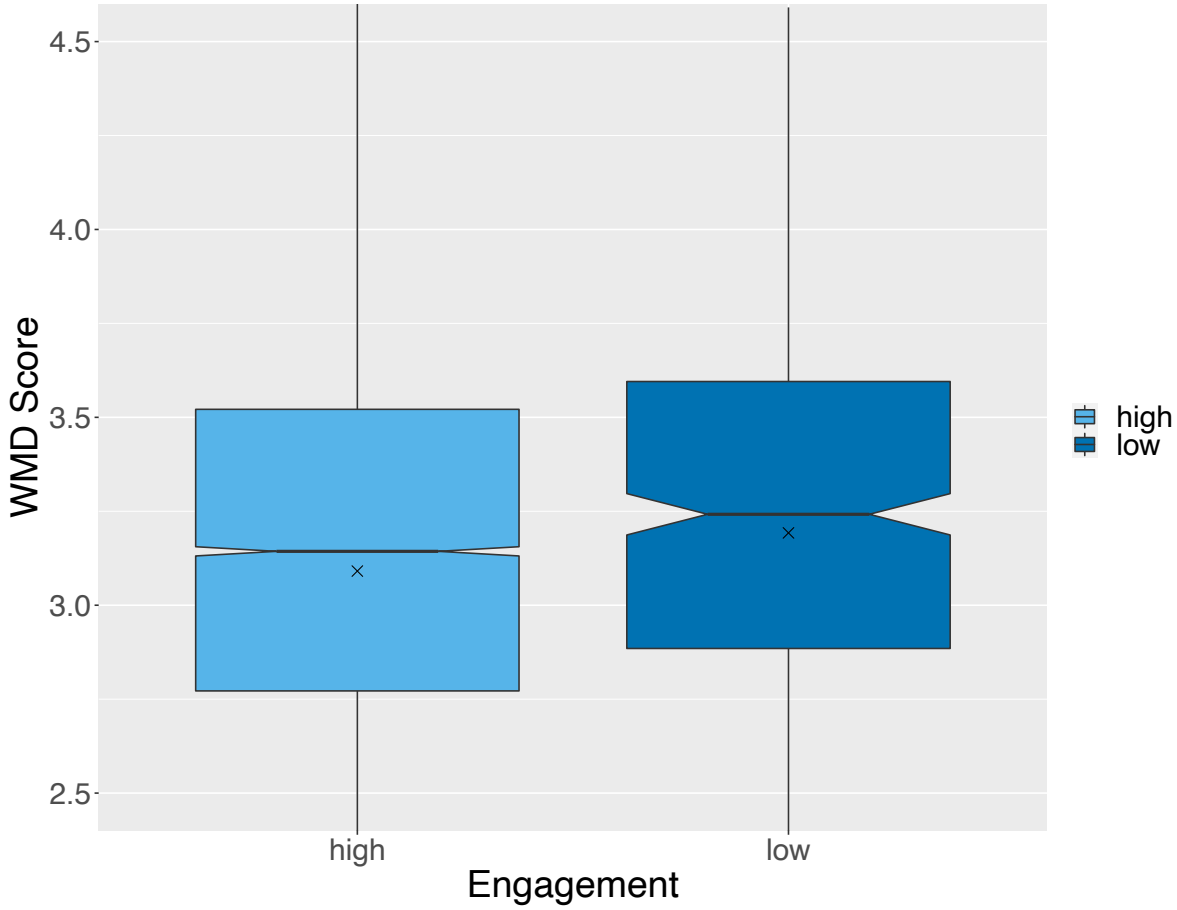


Figure 5.2: WMD Score Distribution by Engagement Annotations

Here, it is noteworthy that spurts from the speakers who were annotated with low engagement have slightly higher mean and median WMD scores, and are centered within a higher WMD score range than spurts from speakers who were annotated with high engagement.

Note that the judgement of speaker engagement is at the level of the task, based on the general impression of the annotator. I believe that these judgements also apply at the spurt level, since it was the cumulative effect of individual speaking turns showing a lack of engagement that gave the annotators the impression that the speaker was not fully engaged. Therefore, from this we can conclude that the WMD score can serve as a measurement of engagement. As for the relationship between WMD score and engagement, lower WMD

scores correspond to higher levels of engagement, however the magnitude of the difference is small.

5.3 WMD Score and Stance

Now that we have established that WMD score is an adequate measurement of engagement, the next set of experiments seeks to determine the correspondence between the strength of stance expressed and the level of engagement.

A Pearson χ^2 test (Pearson, 1900) shows a statistically significant relationship between stance strength and WMD score for the ATAROS 3I and 6B tasks separately ($p < 0.001$) and combined ($p < 0.001$). These measurements do not reach significance for the PSI corpus. This is likely because in the PSI hearings the flow of the conversation, and the instances of stance taking, were not as spontaneous as they would be in natural speech (Haddington, 2004); it is probable that both sides entered into the discourse with some talking points they wanted to address, and although the conversation was overall spontaneous, there was likely some idea of the nature of the questions that would be asked, and some pre-rehearsed answers. For this reason, the choice of words was not influenced or primed by the speaking partner to the extent that it would have been in the ATAROS tasks. Therefore, further experiments will focus on the ATAROS corpora.

Figure 5.3 shows how the range of WMD scores differs between the three stance strengths. Note that for each stance strength, the WMD scores for the 6B task, which was designed to elicit a higher level of engagement, are slightly lower than the 3I task.

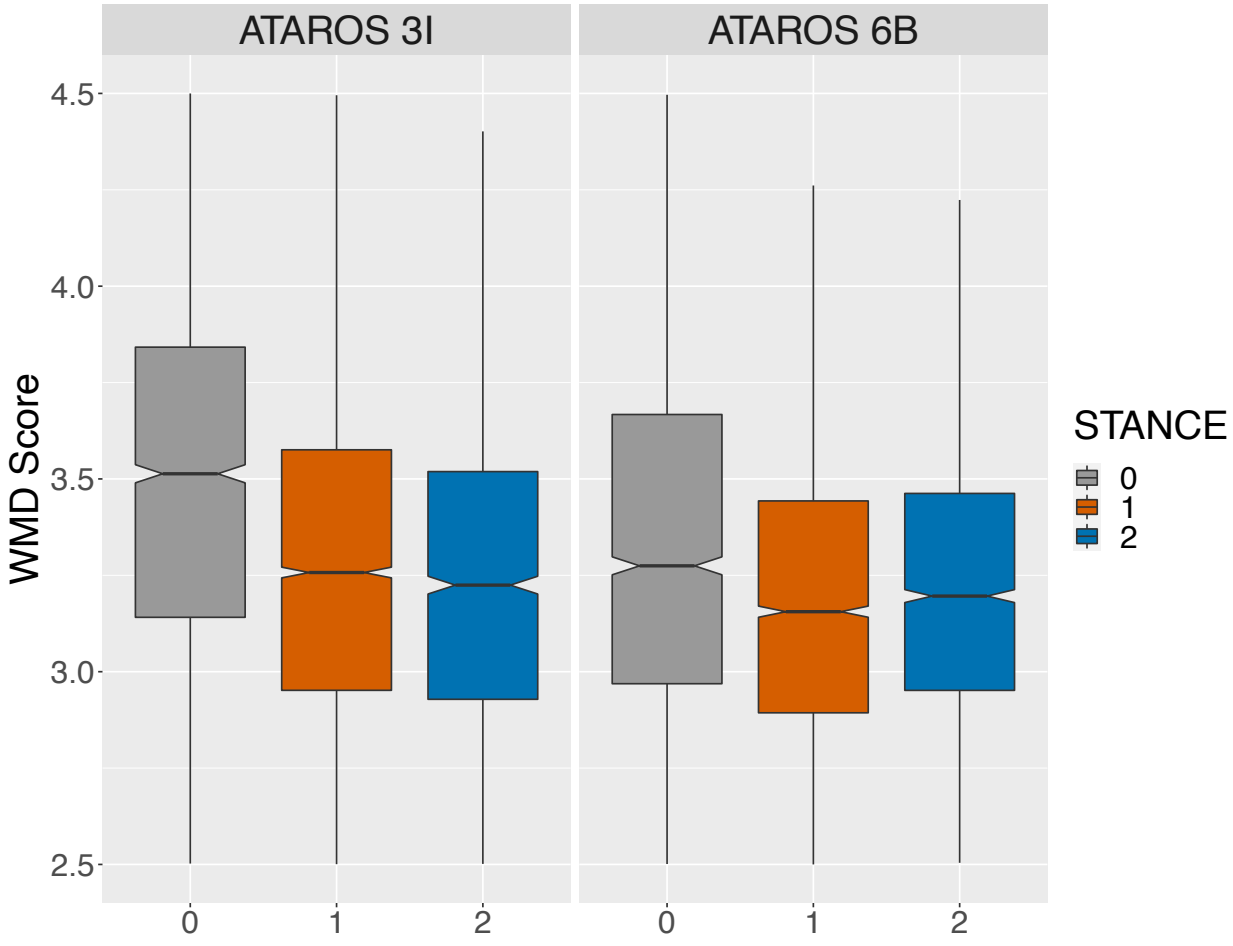


Figure 5.3: WMD Scores by Stance Strength

From here it is clear that the WMD scores for stance strength 0 are higher than those for stance strength 1 and 2 for both tasks. There is little difference in the scores between stance strengths 1 and 2. An ANOVA and Tukey Post-Hoc shows a significant difference between the following stance pairs: 0 - 1 and 0 - 2 for the ATAROS 3I corpus ($p < 0.001$), 0 - 1 and 1 - 2 for the ATAROS 6B corpus ($p < 0.001$), and all three levels of stance strength for a combined 3I and 6B corpus ($p < 0.001$). The ATAROS 3I corpus shows a monotonic decrease in WMD score from stance strength 0 to 2, while the ATAROS 6B shows that stance strength 2 has a range of WMD scores slightly higher than 1.

An investigation into the zero WMD scores, which stood out in Figure 5.1 shows that these scores were not enough to influence the results shown in Figure 5.3. While there is a higher instance of zero WMD scores for stance strength label 0, overall, the proportion across all three stance strength labels is low. This is shown in Figure 5.4. Note that the y-axis only goes to 10%.

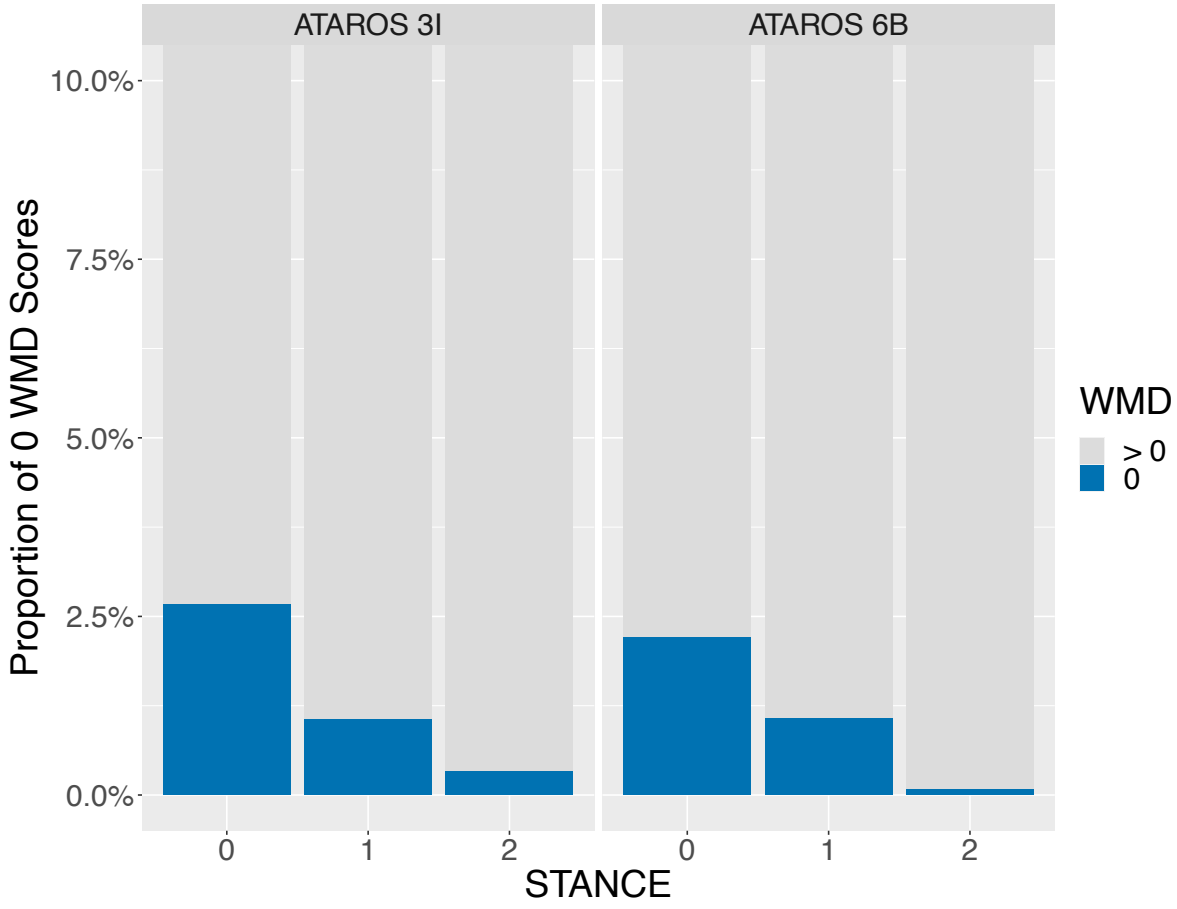


Figure 5.4: Stance Breakdown of 0.0 WMD Scores

Y-Axis is clipped to show detail

From this, the main conclusion is that stance-laden spurts have a higher level of engagement, as reflected in a low WMD score, than stance-less spurts. There is no consistent difference in the WMD scores between strong and weak stance.

As an example to show how WMD scores reflect dialogic resonance, here is an excerpt from the ATAROS 3I task:

A: Kay, books. (0)

B: Books. (0) Um. (0)

A: Don't know if we have anything like books. (1)

B: Useful s- crap. (1) I think, is this - oh, but - what? (x) But these - (x) Huh. (0)

A: Maybe books and - (1)

B: We agree that it's in this one, right? The middle one? (2)

A: Yeah, I agree. (1)

B: Okay. So d- maybe we could put it next to toys! I mean, those a- th- they're for entertainment. So let's do that. (2)

A: Yeah, yeah that's - that sounds reasonable. Yeah, exactly. (1) That's a good way to do it. (1)

Here are some sample combinations of spurts, and the corresponding WMD scores. Recall that the score is assigned to the second spurt in the pair:

A: Kay, books. (0) **B:** Books. (0)	0.0
B: Books. (0). **A:** Don't know if we have anything like books. (1)	2.31
A: Don't know if we have anything like books. (1) **B:** Useful s- crap. (1)	3.07
B: Useful s- crap. (1) **A:** Maybe books and - (1)	3.15
B: We agree that it's in this one, right? The middle one? (2) A: Yeah, I agree. (1)	2.25

The first example, of course, aligns perfectly; *Kay* is a stop word and has therefore been removed prior to calculation. The second pair of spurts share the word *books*, and therefore has a relatively low WMD score. The more interesting cases are the utterances involving the term *Useful s- crap*, the third and fourth examples. As a disfluency, the token *s-* is removed prior to calculating the WMD score. While, in the context of the discourse, the reference makes sense (speaker B is using a categorization of task items as being useful or useless, while *crap* serves as a generic term for an item), the lack of semantic relatedness between the words *crap* and *book* yields a high WMD score (WMD *crap-book* = 3.65). Had speaker B used a neutral term, such as *stuff* (WMD *crap-stuff* = 3.20) or had their utterance used a related term such as *novel* (WMD *book-novel* = 2.55), the resonance between those utterances would have been better reflected in the WMD score. The final pair of spurts, both sharing the term *agree*, again has a relatively low WMD score.

The WMD score serves as a means to quantify the level of engagement where engagement is reflected in the use of similar, or related, terminology. When this measurement is applied to stance strength, it has been shown that WMD scores are lower, indicating a higher degree of engagement, in spurts that express some degree of stance relative to those that do not.

In Chapter 6 I will show that the rate of word overlap between sequential speaking turns is low (> 10% on average across all speaking turns) at all levels of stance strength. Therefore, this parallelism is not achieved solely through repetition.

5.4 Terminology

There are many interesting examples among these parallel structures that demonstrate the ways in which speakers do and do not adopt each others' terminology that deserve attention outside of the WMD score.

5.4.1 ATAROS Corpora

In the ATAROS 3I task, much of the negotiation of terminology was in the naming of the sections of the store. There were common tendencies, such as referring to the areas by the class of item they were placing there, such as food or grocery, hardware, or clothing, or by using the name of a store that was familiar to both participants, such as Home Depot to refer to items one would expect to find in a hardware store. Some participants, however, introduced new terminology into the discourse, often using the suffixes *-y* or *-ish*.

An example of the *-y* suffix is introduced by speaker A in the following. It is immediately adopted by speaker B in their response:

A: Maybe above? Cuz I do- I - I don't know. Yeah. Books of matches. (1)

B: Yeah. Sh- (1)

A: Would you say.. Home Depot stuff, or.. **kitchen-y** stuff? (1)

B: Um. (0) I would say **kitchen-y** stuff, but - (2)

In the ATAROS 6B task, the dialogic resonance and shared conceptualization was shown in the way the participants referred to removing items from the list. The task instructions, as given in Freeman (2015), use the word *cut*, so I consider this the default term:

> You are on the county committee in charge of balancing the budget. Below are the departments that are spending too much. Each department has identified expenses that could be cut to reduce costs.
>
> Your task is to decide which expenses should be cut from each department. To appear fair, you must choose the same number of items from each department. You must discuss each item and reach an agreement about whether to cut it or continue funding it.

Among the alternatives, the most frequent and neutral term was *get rid of*, which was used by 26 out of 29 dyads. Among those dyads that used the term, it was used by both partners in 17 of those dyads, and only one speaker in 9 of them.

Interestingly, of the less frequent alternatives, such as *axe*, *drop*, *nuke*, and *can*, most were used by only one partner. This is likely because these terms were perceived to hold strong negative connotation[1]. Where an alternative term was adopted and used by both speaking partners, the uses tended to appear in adjacent speaking turns, or after repeated uses of the term by one speaking partner.

It seems that, as long as the conceptualization associated with the proposed term does not hold any strong negative meaning, it might be adopted by their speaking partner, if not immediately, then after multiple uses. Further experimentation would be warranted to determine the conditions under which these adoptions happen.

One particularly interesting example of engagement as demonstrated through a shared conceptualization comes in the interaction between two participants in the ATAROS 3I task. They use a re-occurring theme, task items as weaponry, which they refer to when making task-based decisions. These references are made by both speakers; it is not applied to every task item, but it is revisited wherever they can justify using it.

It begins early in the task with speaker A grouping the task items *mousetraps* and *small axes* together under the concept of *things that kill*. Speaker B acknowledges this

[1]This is supported by the scores for these words in the NRC-VAD Lexicon (Mohammad, 2018)

conceptualization and expands upon it in their own use in their use of the term *deadly weapons*.

A: So I'm gonna put it, like - (2) Well.. small axes and mousetraps both kill things. How's this? Is that good? (1)

B: Uh - (0) Yeah, that's true. Any - any - any deadly weapons. Yeah. (1)

A while later, speaker A refers to the task item *half-inch tubing* as "Not quite deadly, but construction-ish." The conceptualization is later expanded to *sharp and potentially deadly things* with the addition of the task item *scissors*. The following exchange shows that this new conceptualization has been accepted by both speakers:

B: Small saws. (0)

A: Sharp objects. Next to the small axes. Yes! (2)

B: Well - (0) Put it with the small axes. Yeah. We don't carry any large weapons. (2)

This dialogue serves as an example that these forms of dialogic resonance need not be limited to adjacent speaking turns, but can re-occur throughout an interaction. This shared conceptualization facilitates their decision making since they can refer to it whenever necessary. It also shows that they are united in how they regard the task itself; they are not taking it too seriously as something that has real-world implications, however they are working toward completing the task and are definitely having fun with it.

5.4.2 PSI Corpus

While the ATAROS tasks were co-operative, the PSI corpus was inherently adversarial. While the spurt-level WMD measurements were not statistically significant, this corpus offers many examples of the speakers' lack of shared conceptualization, which is made evident by

the words they used to refer to common concepts within their interactions. In spite of their lack of conceptual pact, it was clear to both participants what the references were; the lack of agreement on terminology served to emphasize the adversarial nature of the dialogue.

Through multiple passes of the transcript and recordings, I identified several recurring topics and annotated all mentions of them. Any annotations for which I was uncertain were omitted.

As the main reason for the inquiry, the financial crisis was mentioned many times throughout the hearing. Its first direct mentions came in Lloyd Blankfein's opening statement, where he referred to it very neutrally and by its accepted name: *the financial crisis* or simply *the crisis*. While the word *crisis* in isolation holds negative connotation, in this context the term has lot much of its negativity. The subcommittee members used decidedly less neutral terminology ranging from the mild *difficulties* and *problems* to strong terms such as *meltdown* and *debacle*. John McCain referred to it as "the financial crisis that has brought on the .. greatest recession since the Great Depression" and Ted Kaufman continuously referred to it as *this thing*.

One of the more contentious topics was the money given to many major financial institutions, including Goldman Sachs, as part of the 2008 Emergency Economic Stabilization Act[2], commonly known as the bailout. Lloyd Blankfein introduces the topic in his opening statement, where he refers to it as "decisive and necessary government action" and twice refers to it as "an investment from the government." He makes it a point to emphasize that Goldman Sachs paid the loan back with interest, yielding "twenty three percent annualized return for taxpayers." This is in stark contrast with other references to it, primarily by John McCain, where he referred to it repeatedly as "ten million dollars" of the "taxpayers' money." Through referring to it by the amount, and attributing the money to the taxpayers, McCain was strengthening the argument that he was making, which was that Goldman Sachs was not directly affected by the housing market, and therefore did not need the loan.

---

[2] `https://www.congress.gov/110/plaws/publ343/PLAW-110publ343.pdf`, retrieved 11/14/2019

One aspect of the hearings that got a lot of contemporaneous attention in the media was Committee Chair Carl Levin's use of colourful language. It stands out particularly not only because some of it is considered taboo, but also that it breaks the standards of language formality usually observed in government hearings. Lakoff (1989) describes the American legal courtroom environment as "formal and distancing." American senatorial subcommittee hearings are run in a similar manner. Witnesses are called, by subpoena if necessary, sworn in, and asked to provide testimony and undergo questioning by the members of the subcommittee. Witnesses are allowed legal counsel, and all the protections provided by the American legal system.

This formality includes referring to discourse participants by their proper title ("Mister Blankfein", "Chairman Levin", "Senator," et cetera) and many instances of "Thank you Mister Chairman" at the beginning and end of each subcommittee member's turn. Since the purpose of the interaction is to gather information, and in many cases, to get the witness to admit something he would rather not admit, there is little need for rapport building or jocularity. As such, one would expect a formal register of language to be used. Carl Levin breaks that expectation with his use of the informal terms *junk*, *crap*, and *shitty* as in the following exchange:

> CARL LEVIN: (1) I'm deeply troubled by that, (1) and it's made worse .. when .. your own employees (1) believe that those securities (1) are junk (2) or **a piece of crap** (2) or **a shitty deal**, .. words that (2) those emails .. show your employees believe (2) about a number .. of those deals.
>
> (2) Billion-dollar Tim- Timberwolf, .. synthetic CDO squared. .. CDOs get squared now. (2) A senior executive called it **a shitty transaction**, but the Goldman sales force .. was told that it was a priority item for two straight months.

That Levin actually uses the word *shitty* is significant. He could have easily used a much less taboo word to convey the same meaning. Attributing the use of the taboo word to a Goldman Sachs employee not only serves to support his main point, that people within

Goldman Sachs were aware that these CDOs were not stable, but also to discredit Goldman Sachs as a company since the use of such language is commonly associated with a a lack of competence, trusthworthiness, and intelligence (Bradac et al., 1979; Cavazza and Guidetti, 2014; DeFrank and Kahlbaugh, 2019; Rassin and Heijden, 2005)[3]. It also very clearly conveys his opinion of the CDOs.

Levin's use of such language also serves a more personal purpose. Taboo language is almost universally considered a form of intense language (DeFrank and Kahlbaugh, 2019; Jacobi, 2014; Scherer and Sagarin, 2006), and the use of intense language is correlated to perceptions of sincerity (via informality and emotionality) (Bradac et al., 1979; Cavazza and Guidetti, 2014; DeFrank and Kahlbaugh, 2019), believability (Rassin and Heijden, 2005; Slovenko, 1982), and credibility (Scherer and Sagarin, 2006), at least among audiences predisposed to agreeing with the message (Scherer and Sagarin, 2006). Since Carl Levin was the Chairman of the subcommittee, already a well established figure, and it was not in dispute that those particular deals he referred to had failed, he was not likely to experience any negative consequences for having used this style of language. Lloyd Blankfein, on the other hand, could have had an involuntary emotional reaction to the use of such language (Pinker, 2007), particularly in light of the fact that it was attributed to people in his employ. Had he responded emotionally, he might have discredited himself, thus helping to make the case that the investment banks were, at least partially, responsible for the financial crisis.

5.5 Discussion

This chapter investigated ways in which speakers show engagement with their speaking partners and the topic under discussion through the use of parallel linguistic structures, similar terminology, and other demonstrations of dialogic resonance. Through the adoption of each other's language, speakers are showing that they share a broad conceptualization of the concepts under discussion, or of the environment in which the discourse is taking place.

[3]Many studies, however, show the opposite to be true (Cavazza and Guidetti, 2014; DeFrank and Kahlbaugh, 2019; Rassin and Heijden, 2005; Scherer and Sagarin, 2006)

Conversely, refusing to converge on terminology strengthens and emphasizes the disagreement between parties.

The lack of statistical significance for the PSI corpus can be attributed to many possible factors, one of which is the performative nature of the dialogues. Speakers went into the discourse with some pre-planned speaking points that I cannot imagine were unrehearsed. Once the natural flow of the dialogue presented an opportunity to them, the speaker likely fit their speaking point into the flow. This is in contrast to speaking points that are more a direct reaction to what the speaking partner says. Additionally, these speaking turns were longer than those in the ATAROS corpora, with the majority of speaking turns containing multiple spurts. In responding to such a speaking turn, speakers will focus on only one point of many, while in the ATAROS corpus, the back-and-forth nature of the dialogue had speakers addressing a single point with each speaking turn.

This study has shown that the level of engagement, as measured by Word Mover's Distance (Kusner et al., 2015), is greater in stance-laden expressions than in stance-less expressions. This was measured at the level of the spurt because spurts correspond roughly to an intonational phrase (Shriberg et al., 2001), which form coherent thoughts currently active in the mind (Chafe, 1994; Sakita, 2006). Since WMD is a measurement of distributional similarity, this measurement should handle both cases of agreement and disagreement.

As for the relationship between WMD score and stance strength, further study is required. I have shown that WMD scores are lower in stance-laden speaking turns relative to those that do not express any stance, however, this is a static measurement between pairs of utterances, and engagement and dialogic resonance are dynamic processes that develop throughout a discourse. Individual or repeated instances of dialogic resonance, such as the adoption of each other's terms or an ongoing shared conceptualization are only part of the process. These acts do not only happen in stance-laden or stance-less speaking turns, but across all speaking turns.

Chapter 6

LINGUISTIC STYLE COORDINATION

While the previous chapters focused on speakers' use of content words to express varying degrees of emotion, or to align or dis-align with their speaking partner, this chapter focuses on an aspect of language use that we are less conscious of (Niederhoffer and Pennebaker, 2002; Danescu-Niculescu-Mizil et al., 2012), function word use.

In any spoken or written interaction there is some degree of Linguistic Style Matching (LSM) between participants (Niederhoffer and Pennebaker, 2002). This is the conscious and unconscious similarity in language use between speakers. It manifests itself in similar word counts, turn lengths, word use, and even similarities in the classes of words used (Niederhoffer and Pennebaker, 2002). These similarities hold for interactions whether they are face to face or online, between strangers or people who have known each other for a long time, and for "low quality" interactions where the participants did not feel any sense of engagement with the other person (Niederhoffer and Pennebaker, 2002). It also appears to be independent of topic, content or conversation length and is largely unintentional and unconscious (Niederhoffer and Pennebaker, 2002).

Danescu-Niculescu-Mizil et al. (2012) used these patterns of LSM to show power differentials among groups. The specific data used to show this was Wikipedia meta discussion and Supreme Court oral arguments. In both corpora, there are inherent power differences. The Wikipedia discussions involved users, known as *admins*, with access to restricted features; any user, once they have built up enough reputation, could be elected into an admin position by their peers. In the Supreme Court arguments, there are Justices and lawyers; aside from the inherent power differences between these groups, the voting record of each Justice is well known, so the lawyers often enter into the interaction knowing which Justices are likely to

be sympathetic to their case, and which they need to direct their arguments toward.

The measurement used in the Danescu-Niculescu-Mizil et al. (2012) study is called **Linguistic Style Coordination (LSC)**. It measures the degree to which participants of a lower status coordinate their language toward participants of higher status. Specifically, it tracks the use of a set of linguistic marker categories, all of which are function words. Through using function words, they are able to capture linguistic style rather than content, thus devising a domain-independent measurement. Additionally, it removes a level of intentionality, since the selection and processing of function words is largely unconscious (Ireland et al., 2011; Niederhoffer and Pennebaker, 2002). The specific markers used, as categorized in the 2007 version of LIWC (Language Inquiry and Word Count) (Pennebaker et al., 2007) are: articles, auxiliary verbs, conjunctions, adverbs, impersonal pronouns, personal pronouns, prepositions, and quantifiers. For each marker category, the degree of coordination for a given speaker is the difference between the probability of its use conditioned on previous use and the speaker's baseline use of the marker. Scores are aggregated across markers and group members to calculate the degree of coordination between groups.

In my own study, I am not studying power differentials, but rather stance strength-based differences. I hope to establish whether the rate of linguistic style coordination differs based on the strength of the stance being expressed. Since the participants in the ATAROS corpus are matched for age, and the balance of matched and mixed gender dyads controlled for, I do not expect to see any overall effect of power differential; the PSI corpus, on the other hand, has an inherent power differential. Additionally, rather than aggregating over a group, I will be calculating a per-participant score for each marker and each stance strength label. Overall, I will be following the procedure outlined in Danescu-Niculescu-Mizil et al. (2012), with the following changes: the equation will reflect stance differences rather than group membership, and I will be using the dictionary from the latest version of LIWC (Pennebaker et al., 2015). Following the results of the Danescu-Niculescu-Mizil et al. (2012) paper, and assuming some correlation between strong stance and power (that the person expressing it will be exerting some power, at least temporarily), I predict a lower rate of LSC among

speaking turns that express stronger stance than those that express weak or no stance.

Equation 6.1 shows the calculation I will be using to compute the rate of linguistic style coordination for an individual speaker. Values referring to the conditioning speaking turn, that is, the immediately previous turn, are indicated with *prev*, M represents a specific marker from the set given above, and S represents stance strength (0, 1, or 2):

$$LSC = P(M|M_{prev}, S) - P(M|S) \tag{6.1}$$

In simple terms, this is the probability of marker use conditioned on the use of the marker in the immediately previous turn and the strength of the stance being expressed minus the speaker's baseline use of that marker at that stance strength.

6.1 Experimental Design

For this set of experiments, the unit of measurement is the speaking turn. A speaking turn is formed out of sequential, uninterrupted spurts by a single speaker. The stance strength assigned to the speaking turn is the maximum score assigned to its component spurts. The reason for using the speaking turn rather than the spurt is that while not everything one says needs to relate directly to what was said immediately preceding, in forming an utterance, a speaker is definitely influenced by the most recent speaking turn, if only subconsciously.

Words are case normalized. The LIWC dictionary was designed to analyze text from the internet and social media, and includes the same variant spellings used by the ATAROS transcribers, so tokenization and lemmatization are not required. Each word in a speaking turn is queried against the LIWC dictionary; if the word is found, the counter for all marker categories in which it is found are incremented[1]. The marker categories used in this study, following Danescu-Niculescu-Mizil et al. (2012) are: articles, auxiliary verbs, conjunctions, adverbs, impersonal pronouns, personal pronouns, prepositions, and quantifiers.

Counts are stored on a speaker by speaker basis. For the purposes of calculation, each

[1]Note that this is a string match, and will also increment homonyms.

task is considered a separate corpus, therefore, the ATAROS participants appear in two corpora. The probability calculation, given in Equation 6.1, is calculated over the entire task. Where a specific marker is not used by a speaker, the resulting LSC score is 0; where their speaking partner does not use a marker, the LSC score assigned is also 0, since the true value of the equation is undefined.

Since speaking partners do not necessarily align along all markers, nor even, reciprocally, across the same marker (Danescu-Niculescu-Mizil et al., 2012), I am also calculating an aggregate score over all markers; this is the average of all the non-zero individual marker LSC scores.

Figures 6.1, 6.2, and 6.3 show histograms of the distribution of LSC scores across all markers. Recall that a separate score is calculated per-speaker for each marker and that a separate tabulation was made for each stance strength. The Aggregate score is the per-speaker average across all non-zero LSC scores. The histograms are clipped to show the differences in the middle of the range; the upper range of the zero LSC score counts varies from 25 – 85; given the pool of 56 speakers and 3 stance strengths, the entire set of LSC scores for each ATAROS corpus is 168 data points.

The LSC score represents the difference between the probability of marker use conditioned on previous use, and the speaker's baseline probability of using that marker. These figures show scores of zero and above, an indication that there is some conditioning effect of marker use by the speaking partner. There are two scenarios for a zero value. The first is lack of use of a marker by a speaker, making both the conditional and the baseline probabilities zero. The second is that their use of the marker is not conditioned on a previous use, and the conditional and baseline probabilities are equivalent.

These histograms show no obvious difference in the distribution of the LSC scores between the ATAROS corpora; the PSI corpus has only seven speakers, so it is to be expected that the counts will be low. Comparing within-corpus marker use, there looks to be no obvious trend in the range of LSC scores, other than the fact that *Quantifier* scores are the most sparse.

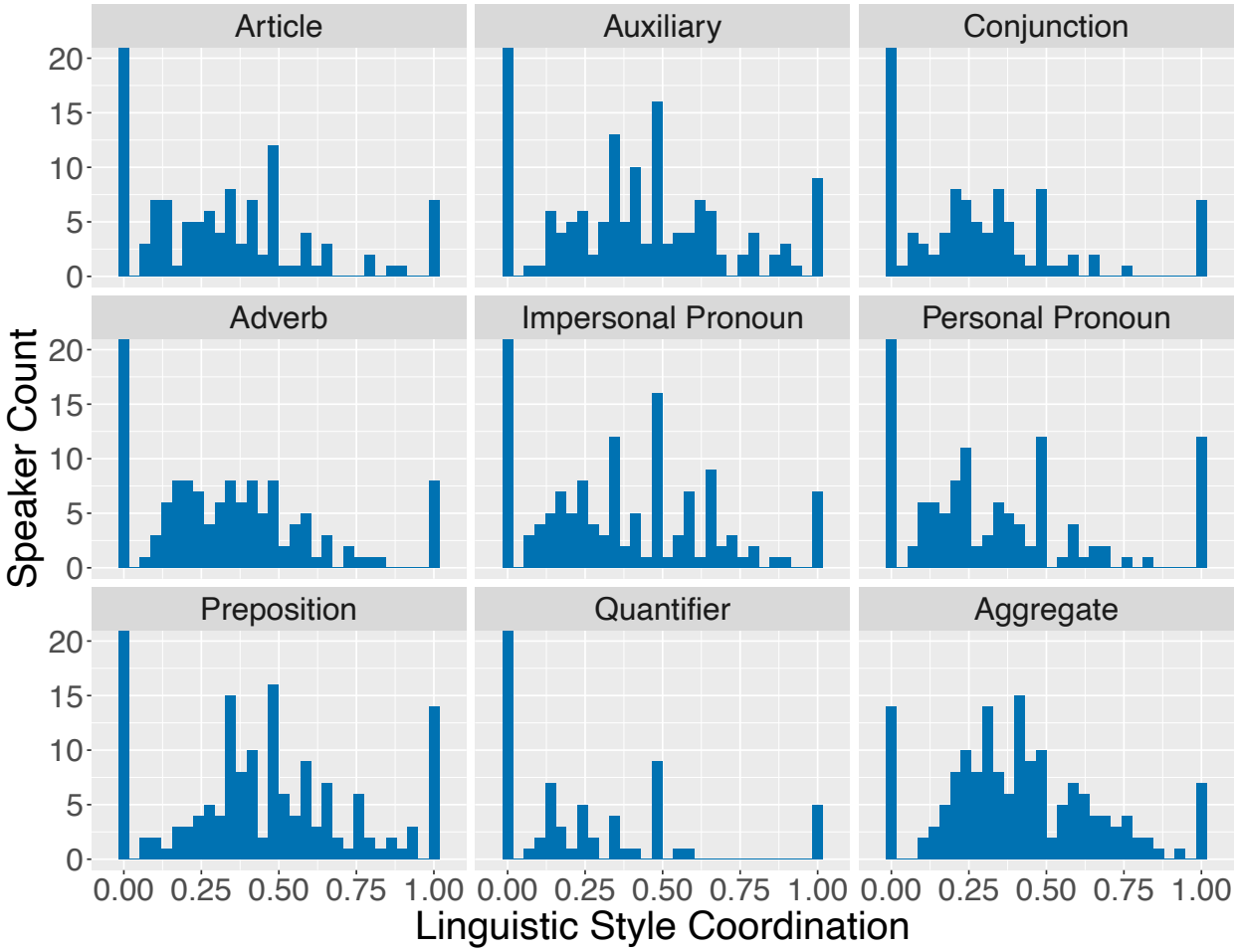


Figure 6.1: LSC Score Distribution Across Speakers: ATAROS 3I

The Y-Axis is clipped to show detail in the middle of the range.

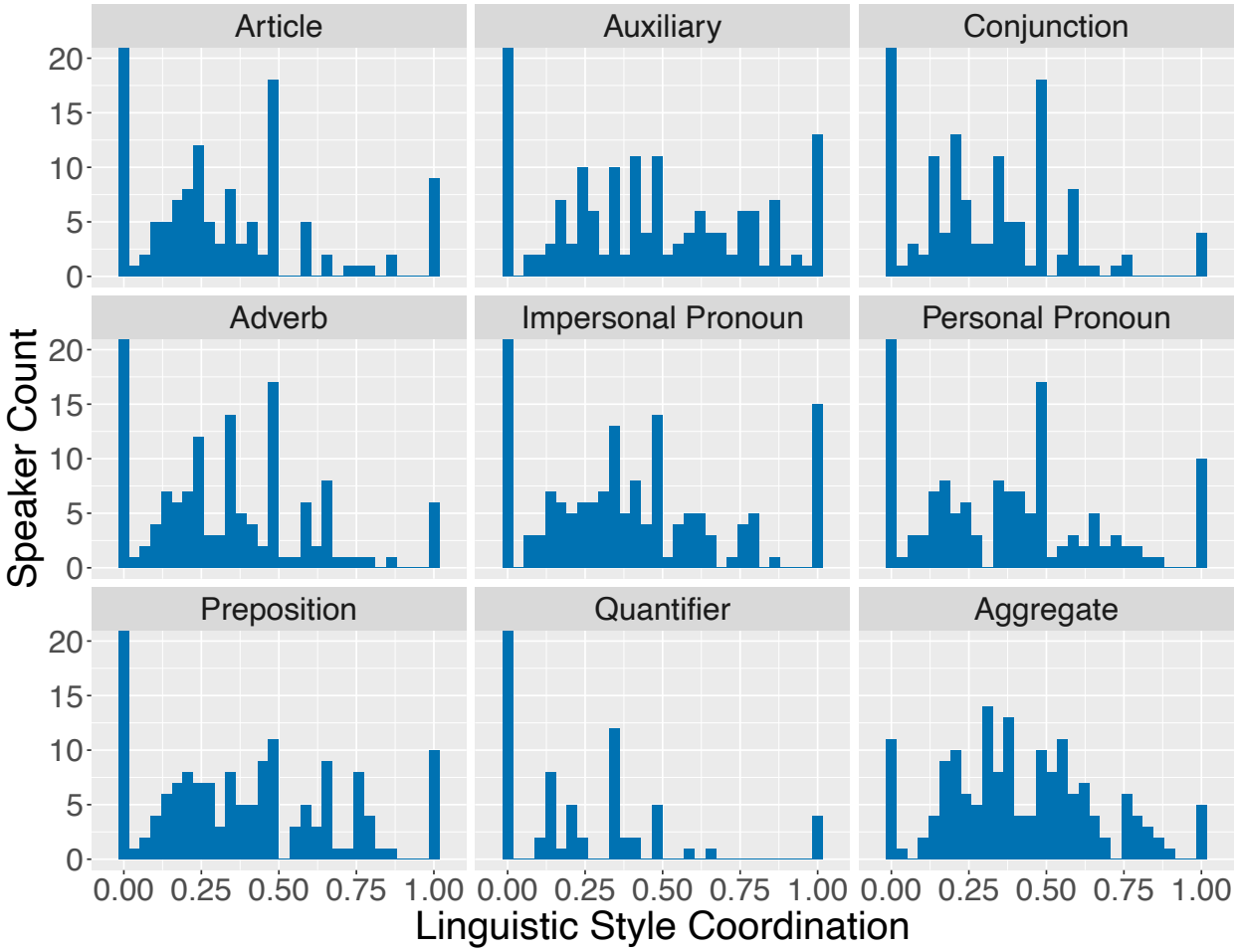


Figure 6.2: LSC Score Distribution Across Speakers: ATAROS 6B

The Y-Axis is clipped to show detail in the middle of the range.

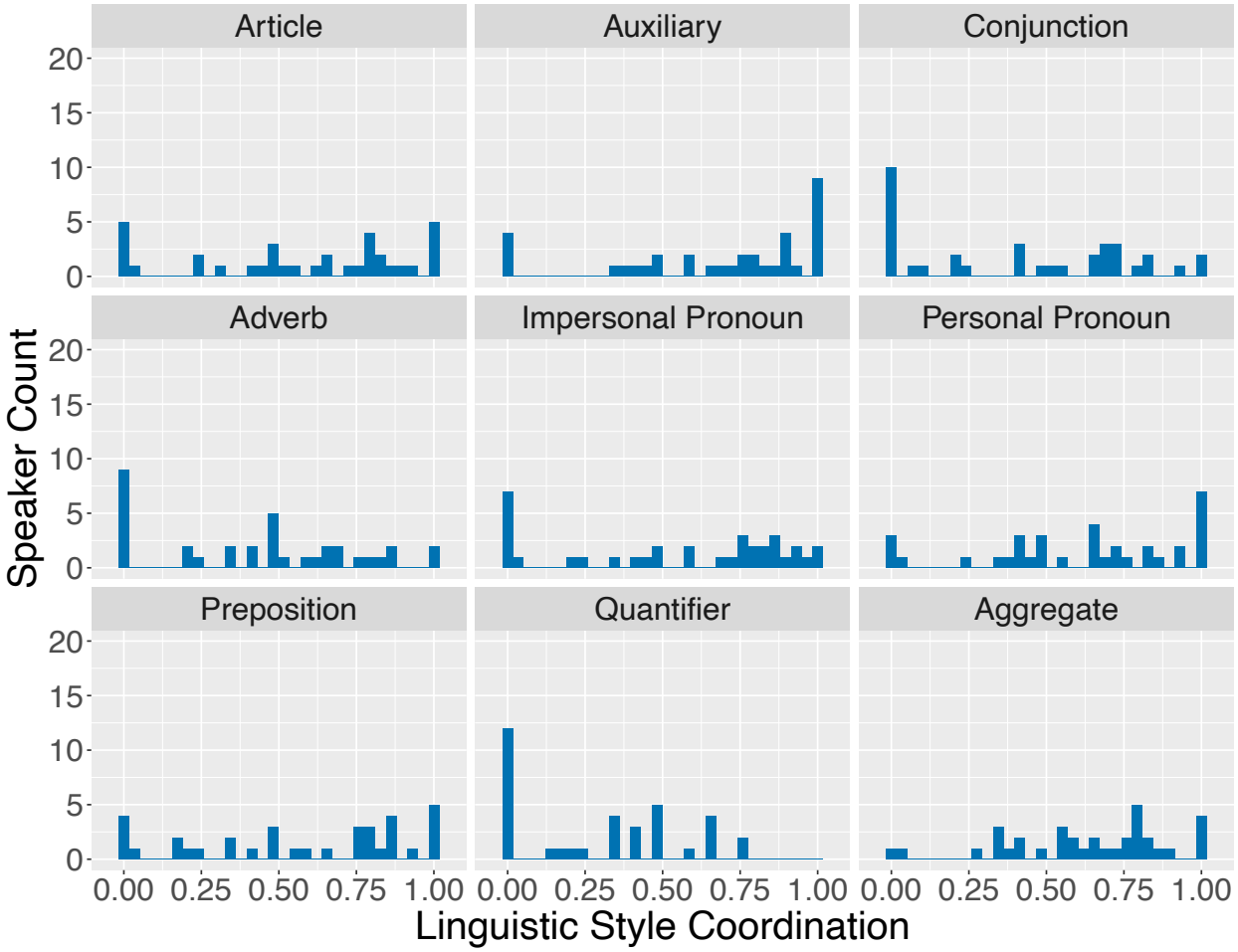
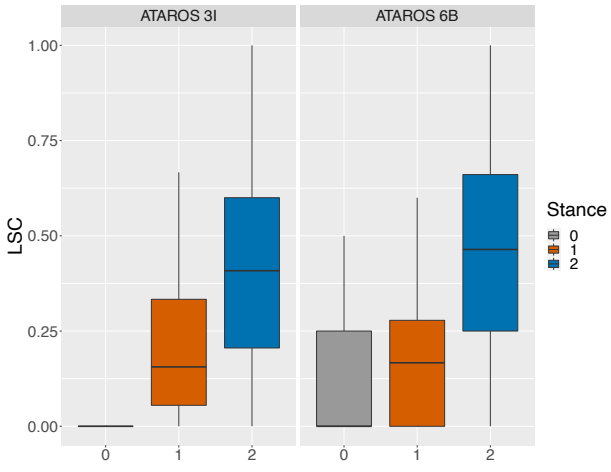


Figure 6.3: LSC Score Distribution Across Speakers: PSI

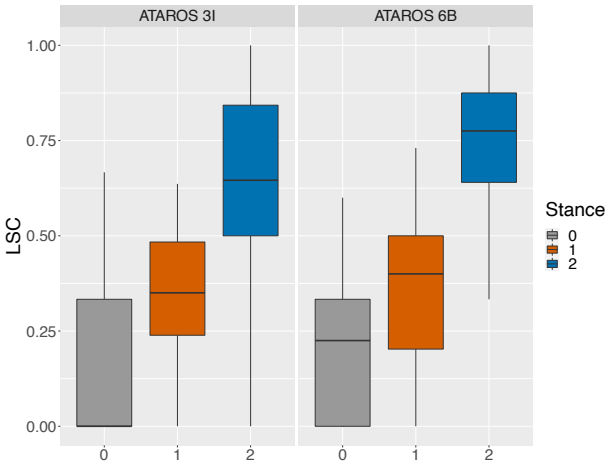
6.2 Linguistic Style Coordination and Stance

A Pearson χ^2 test (Pearson, 1900) shows a statistically significant relationship between stance strength and Linguistic Style Coordination and all markers except *Quantifier* for the ATAROS 3I task ($p < 0.005$), all markers for the ATAROS 6B task ($p < 0.05$), and no markers for the PSI task. The aggregated score does not show significance for any corpus.

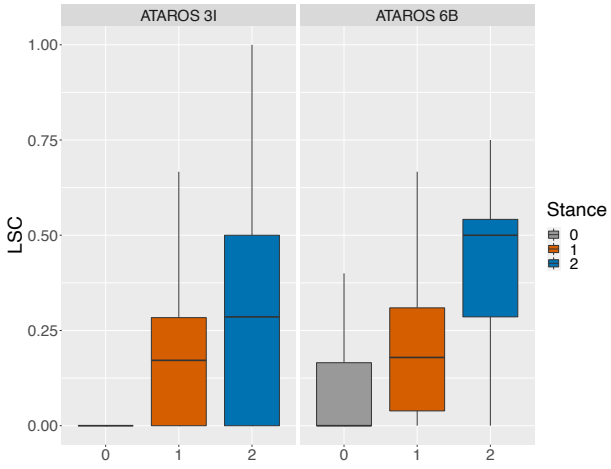
Figure 6.4 shows how the level of LSC differs by stance strength for each marker.



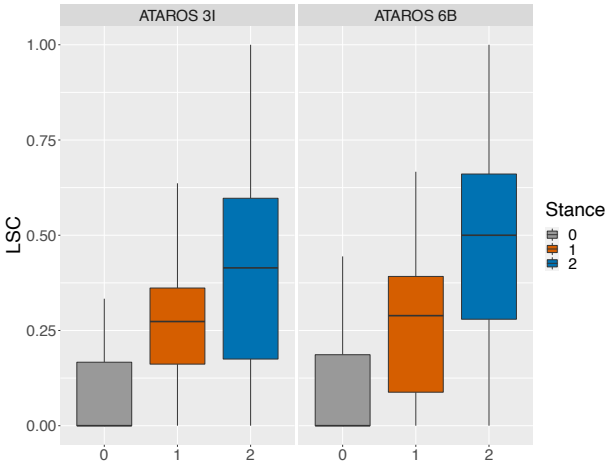
(a) Article



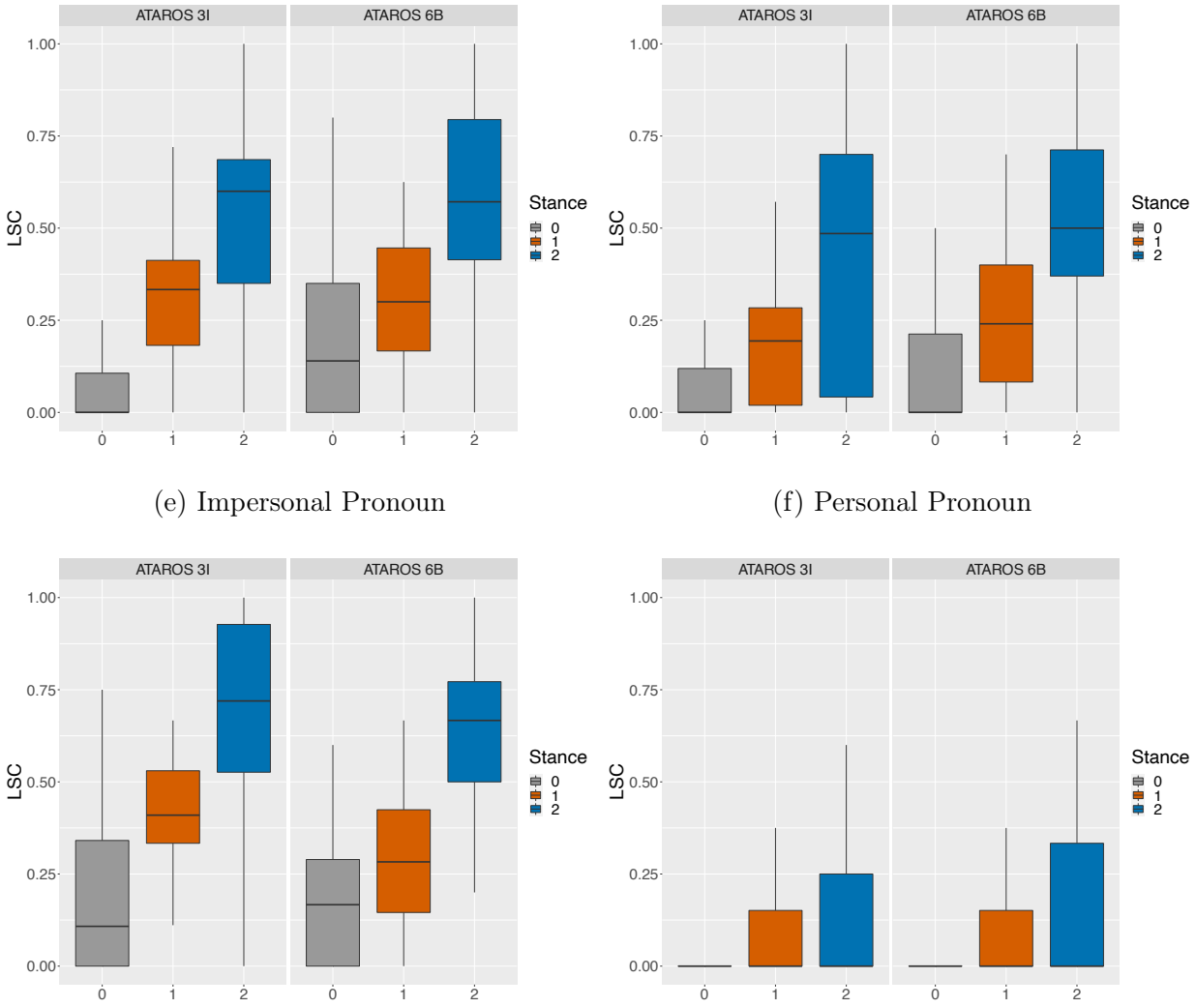
(b) Auxiliary Verb



(c) Conjunction



(d) Adverb



(e) Impersonal Pronoun

(f) Personal Pronoun

(g) Preposition

(h) Quantifier

Figure 6.4: Linguistic Style Coordination Across Markers

Two things stand out from this. The first is the low level of LSC at stance strength 0 overall, and the complete lack of LSC for some markers. This is counter to my prediction that lower levels of LSC would indicate strong stance. This was based on an analogy to the power differentials proven by Danescu-Niculescu-Mizil et al. (2012); I predicated that the expression of strong stance would be similar to a power differential, albeit temporarily, in

favour of the person expressing the strong stance. It is exactly the opposite. Stance strength is positively correlated with LSC.

Recall that there could be two reasons for a zero LSC score: either the speaker or their partner showed no use of the marker, or a speaker's use of the marker is not conditioned on the use of the marker by their partner and the conditional and baseline probabilities are equivalent. An investigation into this shows that only a few speakers showed a complete lack of use for a specific marker at a specific stance strength, only two speakers in the ATAROS 3I task, and three in ATAROS 6B. There was no consistency in which marker was not used, however all instances were at stance strength level 2. This is most likely due to the fact that stance strength label 2 is the least frequent stance annotation, rather than any pattern of marker use. As for equivalent conditional and baseline probabilities, no speaker demonstrated a complete lack of conditioned use for any marker, and only two speakers had an LSC score of less than 0.05 for any marker, both for stance strength label 1. From this, we can conclude that these instances are not because of a lack of conditioning at this stance strength, but that the rate of LSC at stance strength 0 is legitimately low.

The second obvious result is the high level of LSC for stance strength 2. All markers show a much higher range of LSC scores for stance strength 2 than 1. This indicates that, when expressing strong stance, a speaker is more likely to use the same marker categories as their speaking partner, than when expressing weak or no stance.

From this, we can conclude that there is a higher rate of Linguistic Style Matching for strong stance; stance-less speaking turns show a very low rate of LSC, while weak stance is intermediate.

6.3 Discussion

The results in this chapter show that the degree of Linguistic Style Coordination is positively correlated with stance strength. What is not clear is whether this is specifically because of the strength of the stance being expressed, or because of another factor that also varies by stance strength.

It has been a common feature in these ATAROS corpora that turn length, as measured in word count, is positively correlated with stance strength ($p < 0.001$). It stands to reason that longer speaking turns would include more words, and therefore the need for more grammatical structure. In subtracting the speaker's baseline probability of marker use for each stance strength, however, speaking turn length is controlled for in the equation.

Another possibility is the fact that strong stance simply includes more grammatical markers. Figure 6.5 shows this to be the case. Here, only the use of a marker category is being counted, not the total number of markers in the speaking turn, but it is clear that there are more marker types being used in strong stance than in weak or no stance. Again, however, the calculation controls for this by subtracting each speaker's baseline use. The low levels of marker-type variety in stance-less speaking turns, however, at least partially explains the lack of LSC for stance strength label 0.

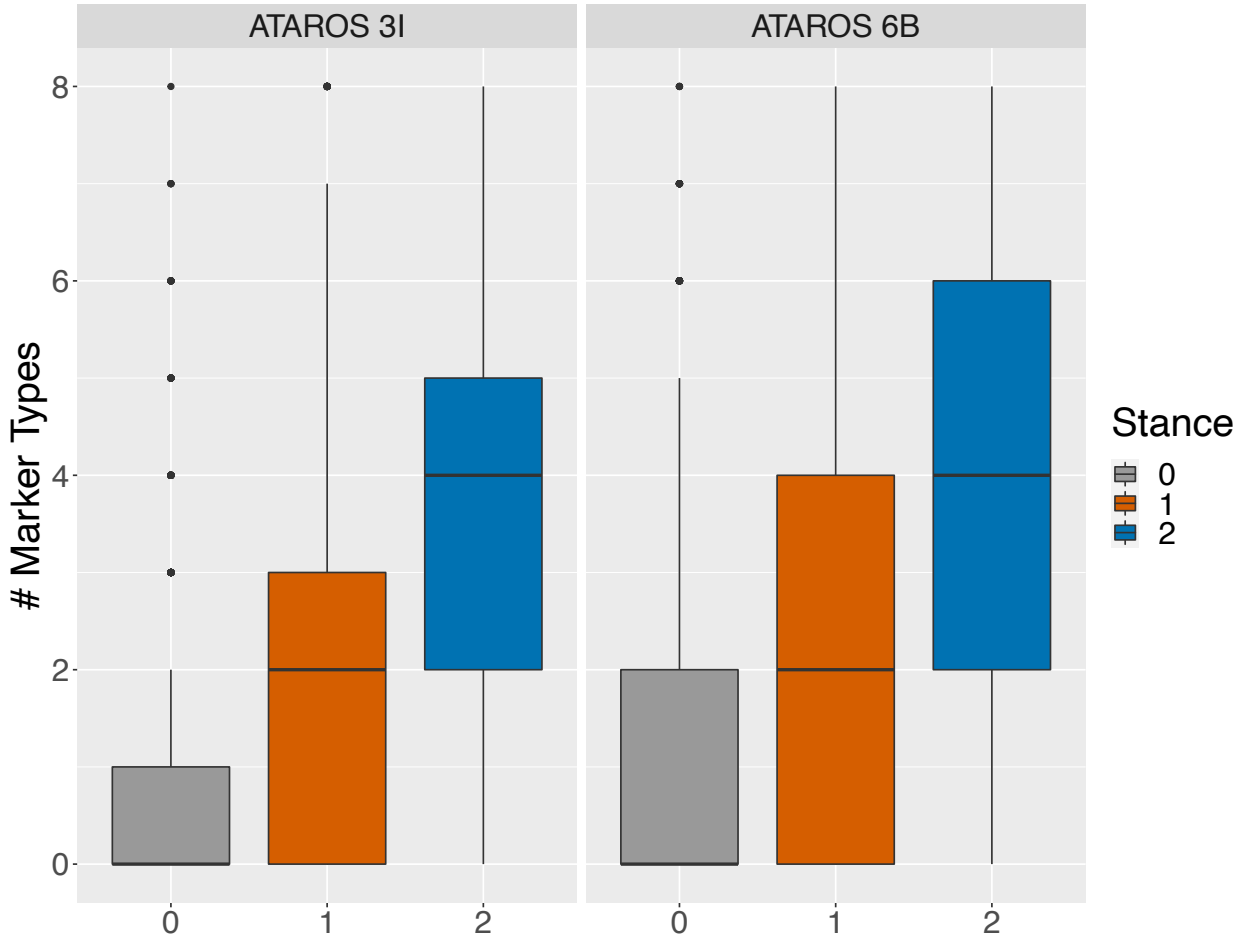


Figure 6.5: Marker Distribution by Stance Strength ($p < 0.001$)

One potentially interfering factor remains. Chapter 5 shows that stance-laden speaking turns show a higher rate of terminological alignment, as measured in Word Movers Distance (WMD) (Kusner et al., 2015), than stance-less speaking turns. Therefore, it is possible that the overlap in marker use between sequential speaking turns is due to dialogic resonance in the form of parallel sentence structure. Figure 6.6 compares this overlap. *% Olap* represents the proportion of words in the current speaking turn that appeared in the immediately previous speaking turn; *% Markers* represents the proportion of overlapping words in the current speaking turn that are markers. The relationship between proportion and stance

strength is significant ($p < 0.001$). Note that this measurement uses string matching rather than lemmatization, and was therefore unable to match morphological variants of a word. With these limitations, the proportions are overall very low, though stance strength 2 shows a higher level than 0 or 1.

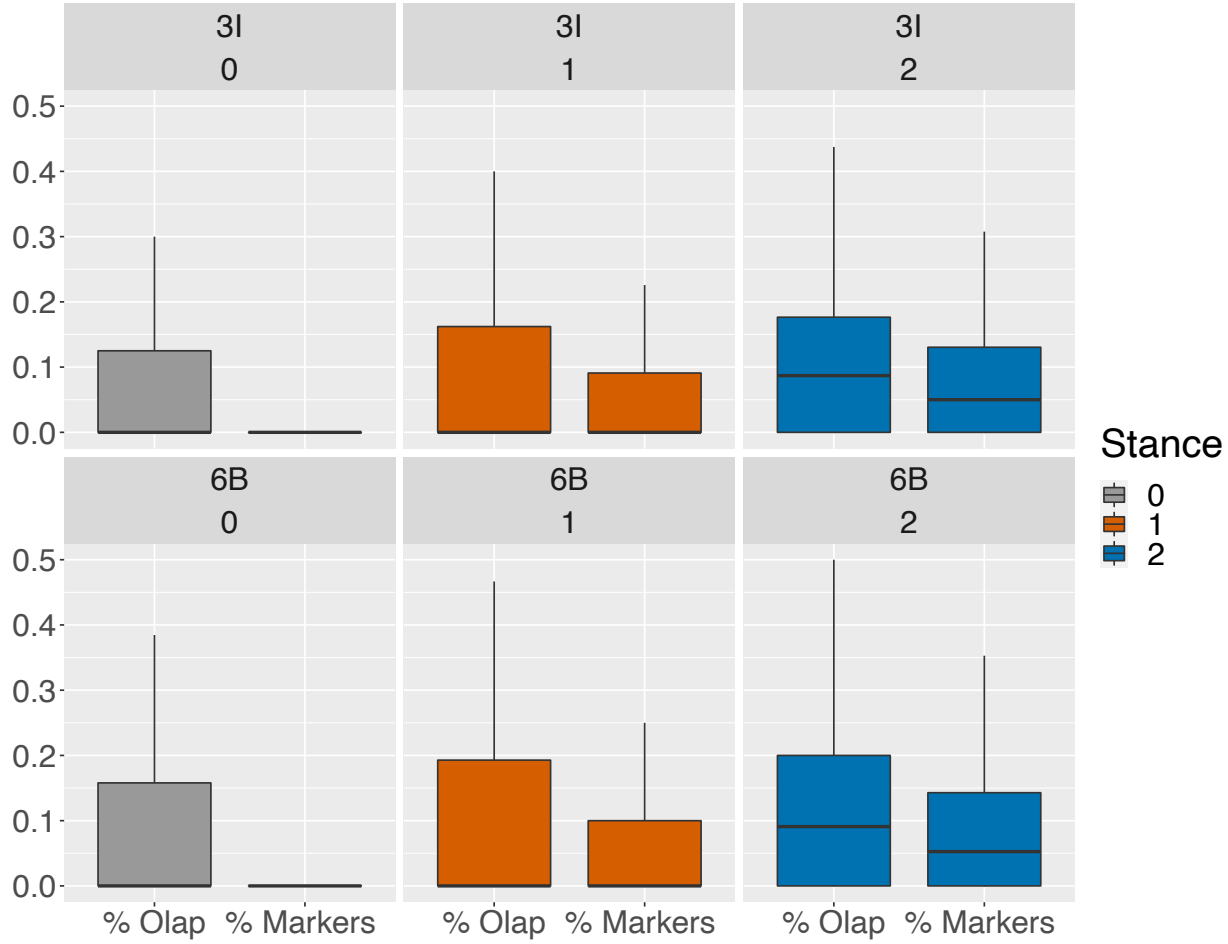


Figure 6.6: Proportion of Words Shared Between Previous and Current Speaking Turn

This lessens the likelihood that dialogic resonance is a factor influencing the LSC results. The higher levels of Linguistic Style Coordination in expressions of strong stance are most likely related to the strength of stance being expressed rather than other factors. There

are several reasons that could explain this. One is that expressions of strong stance may be more complete or coherent in the mind of the speaker; the speaker is likely to show more confidence in strong stance than weak. The rate of marker-type use in strong stance might speak to more complete, less fragmented sentence structure. A second possible factor might be a tendency for speakers to support their statements of strong stance with evidence or justification. Finally, given the laboratory setting of these corpora, the speakers might hedge their statements of strong stance to mitigate any sense of dominance or power they feel they are exerting over their speaking partners. Any of these could add to the length of the speaking turn, and frequently employ closed-class words such as modal and auxiliary verbs, conjunctions, prepositions, et cetera, all of which are included in our marker categories. Further study would be warranted to determine the reason for this behaviour.

This chapter has shown that, even along a dimension as unconscious as the coordination of function word categories (Danescu-Niculescu-Mizil et al., 2012), speakers show different patterns when expressing stance compared to when they are not and, for many markers, differences between strong and weak stance. While this could be partially attributed to an increase in marker use among speaking turns that express strong stance, there is some indication that the increased coordination in function word use corresponds to the strength of the stance being expressed.

Chapter 7

CONCLUSION

This study is one of the first computational studies to investigate dialogical aspects of stance taking in spontaneous spoken English dialogue with a focus on the lexical choices made by speakers and how speaker behaviour differs depending on the strength of the stance being expressed. I focus on dialogical aspects since this has not been addressed in computational studies to date though it has been well established that speakers influence each others' lexical choices and aspects of their grammatical style in any dialogic interaction (Brennan, 1996; Niederhoffer and Pennebaker, 2002). When comparing the acoustic-prosodic markers of stance taking found by Freeman (2015, 2019) to speaking style and unigram features, including punctuation, Levow et al. (2014) found that the unigram features, among them several that were not overtly associated with stance taking, to be the most discriminative in a stance detection task. This shows that there are subtle signals of stance taking in dialogical interactions outside of the overt use stance-laden lexical or grammatical structures.

Though surrounding context was available to the annotators, stance strength annotations were assigned at the level of the spurt. Therefore all statistical analysis and classification studies using the ATAROS corpus were also at this level. Additionally, most computational research on stance or subjectivity detection to date has focused on single threaded units, such as news articles or reviews; even when there was some dialogical aspect to the content, such as forum posts, or debates, focus still remained on the content of the turn in isolation. By investigating dialogical factors, I am able to study how context outside of the current speaking turn may affect the words chosen to express stance.

First I looked at the strength of words along the emotional dimensions of *valence*, *arousal*, and *dominance*. I tested whether the strength of the strongest word in the spurt was corre-

lated with the perceived strength of the stance being expressed. A regression model showed that the probability of each stance strength correlated with the dimensional strength along all three dimensions; weaker stance was most likely at the lower end of the score range (weak emotive strength, not excited, and submissive), and strong stance was most likely at the upper range (strongly emotive, aroused, dominant).

I then looked into whether different stance strengths were conditioned on different ranges of these dimensional scores, and found that consistent patterns emerged only in the relationship between these word scores and strong stance. I showed that strong words, at both extremities of the valence scale were strongly associated with strong stance. As for the dimensions of arousal and dominance, there are similar findings. High arousal and dominance scores are also associated with strong stance.[1]

Next, I looked at a measurement of dialogic engagement and resonance as demonstrated by the use of similar terminology. The measurement of dialogic engagement was spurt-level Word Mover's Distance (WMD) (Kusner et al., 2015), which was calculated pair-wise among the spurts in adjacent speaking turns. Lower scores represent more lexical or semantic similarity. This measurement was shown to correspond with human annotated judgements of engagement. I found a significant relationship between WMD score and stance; higher WMD scores were found in stance-less spurts than in stance-laden spurts, but there was not a consistent nor noticeable difference between strong and weak stance.

Finally, I looked at stance-based differences in the rate of Linguistic Style Coordination (LSC) (Danescu-Niculescu-Mizil et al., 2012), that reflects dyadic similarities in the use of function words. It is a conditional probability calculation, calculated over the entire task. I found that speakers showed a higher rate of LSC in turns that express strong stance than those that express weak or no stance.

While none of these factors in isolation have been tested in a classifier as strong signals of stance taking, the following section will show how these factors correspond to the results of Levow et al. (2014).

---

[1]Recall that the dominance and arousal scores are are on a bipolar [0 – 1] scale, where 1 represents the "excited" end of the arousal scale, and the "in control" end of the dominance scale, while strength on valence scale represents the distance from the neutral point.

7.1 Correspondence to Other Classification Results

Recall that Levow et al. (2014) compared acoustic prosodic features to speaking style and unigram features, and found that the unigram features, which includes punctuation, were the most discriminative in a stance strength classification task. The specific set of unigrams is given again in Table 7.1.

yeah	okay	um	hm
need	maybe	important	good
the	this	could	but
?	,	!	

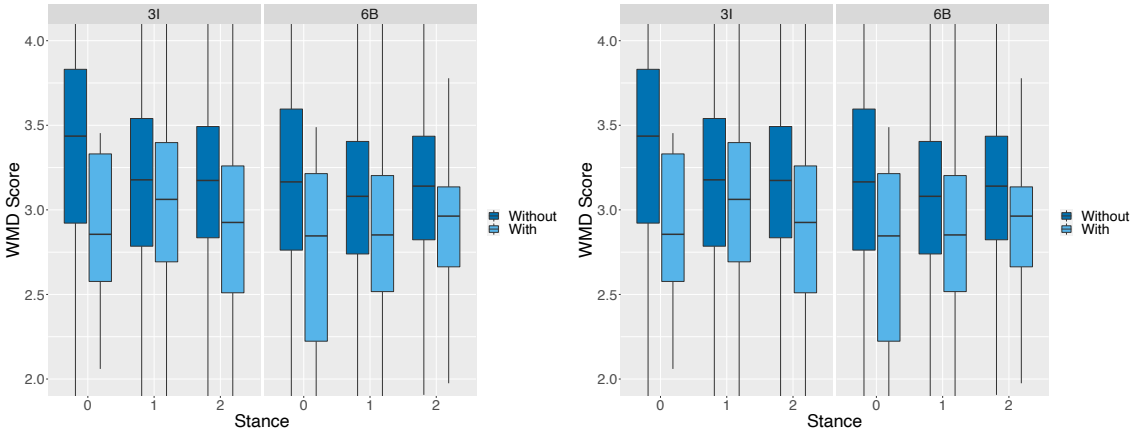
Table 7.1: Discriminative Unigrams Found in Levow et al. (2014)

For some of these unigrams, their relationship to stance taking is obvious. For example *need*, *important*, *good*, and *could* are found in both the Arguing Lexicon (Somasundaran et al., 2007) and Subjectivity Clues Lexicon (Wilson et al., 2005a) while *maybe* is found only in the latter.

Looking at the scores for these unigrams in the NRC-VAD Lexicon, *need*, *important*, and *good* demonstrate a very clear relationship between the score on at least one emotional dimension, the breakdown of the stance strength annotations among spurts containing these unigrams in these corpora, and the predicted stance strength values predicted from the models in Chapter 4. It is clear that these unigrams were the stance-holding lexical item in the majority of spurts in which they appear.

The unigrams *could* and *maybe* do not show this correspondence in NRC-VAD Lexicon scores, however, when comparing the Word Movers Distance (WMD) scores among spurts that contain these unigrams to spurts that do not, spurts that contain these unigrams have lower WMD scores ($p < 0.001$) across all stance strengths. Recall that lower WMD scores represent a greater level of lexical or semantic relatedness. This is shown in Figure 7.1, and

supports the results found in Chapter 5 that stated that spurts expressing stance had lower overall WMD scores than spurts that did not.



(a) maybe, n = 587

(b) could, n = 510

Figure 7.1: Distribution of WMD Scores for Spurts Containing the Unigrams *maybe* and *could*

This shows that, in addition to being signals of stance taking themselves, these unigrams also tended to appear in speaking contexts where there was terminological or contextual overlap with the previous speaking turn. Given the semantics of both words, it is likely they were used in making suggestions or proposals.

As for the other unigrams, their connection with stance taking is a little less direct. Careful analysis shows that the unigrams *yeah*, *ok*, and *um*, show a trend similar to the unigrams *could* and *maybe* in that they occur in spurts with low WMD scores. It is likely that they are also used in making suggestions or expressing agreement, and therefore have some conceptual or terminological overlap with the previous speaking turn. The unigram *hm* is unique in that its appearance indicates a lack of stance, stance strength 0. This is also consistent with the WMD results; spurts that contain the unigram *hm* have higher WMD scores than spurts that no not.

As for the relationship between these unigrams and the Linguistic Style Coordination (LSC) results given in Chapter 6, it should be noted that, since LSC focuses on function word categories, only the unigrams *the*, *this*, and *but* overlap with the discriminative unigrams, belonging to the Article and Conjunction categories respectively. That they are associated with stance-laden spurts aligns with the results of that chapter, which states that there is a higher rate of LSC, showing more coordination of function word category use, in expressions of strong stance than in those expressing weak or no stance.

The final three unigrams, being punctuation marks, show no relationship to the dialogical results found in this study.

While I was not able to validate that the clues of stance taking found in this study would work as individual or combined features in a classification task, due to factors outside the scope of this study, I was able to show how these results correspond to, and can help explain, the results of an existing classification study. It is my hope that these clues can be applied to existing classification systems, most of which were developed to identify instances of stance taking in single threaded documents, so that they can easily apply to stance taking in a spoken, dialogical setting.

7.2 Future Work

In this study, I showed how speaking partners converge lexically and semantically when expressing stance using a pairwise spurt-level measurement, Word Movers Distance (WMD) (Kusner et al., 2015) as a proxy for syntactic and semantic parallelism, and how they converge stylistically among stance-laden speaking turns using Linguistic Style Coordination (LSC) (Danescu-Niculescu-Mizil et al., 2012), a measurement of function word category use calculated over the entire dialogue. I speculate that investigations using actual syntactic parallel structures would yield similar or even stronger results since the WMD measurement, being focused on semantic and distributional similarity, missed some obviously parallel structures. Similarly, a localized measure of Linguistic Style Coordination may reveal subtleties missed by a dialogue-wide measurement.

In listening to the recordings for this study, I noticed many instances of overlapping speech and irregular pauses. These, too, might reveal a subtle signal of stance taking in a dialogical environment. Additionally, since it is a recorded corpus, there may be stance-related clues in patterns of phonetic and prosodic entrainment. Both of these topics would be fertile ground for future research.

Finally, this work focused on two distinct corpora, the ATAROS corpus, which was cooperative and collaborative in nature, and the PSI corpus which was adversarial. Many of the dialogic results I found in the ATAROS corpus did not show statistical significance in the PSI corpus. I speculated that this was due to the differing natures of the interactions; the ATAROS corpus was very interactive, with each speaker contributing shorter, more frequent speaking turns, while the PSI corpus had opening and closing remarks, and longer overall speaking turns. It is possible that the adversarial nature of the interactions in the PSI corpus also had an effect. It would be an interesting exercise to try to replicate these results on an adversarial corpus with shorter, more interactive speaking turns.

7.3 Final Remarks

This study is one of the first computational studies to investigate dialogical aspects of stance taking in spontaneous spoken dialogue, with a focus on the lexical choices speakers make in expressing stance of varying levels of strength. It has shown that there are differences in speakers' dialogical behaviour depending on whether they are expressing stance or not. Speakers show a higher level of dialogic engagement, as demonstrated by the use of similar terminology and parallel linguistic structures, when expressing stance relative to when they are not. When measuring the convergence in linguistic style through function word category matching, it is also higher in speaking turns that are expressing strong stance than it is in turns that express weak or no stance. It is my hope that these findings can be applied to the task of subjectivity detection and classification, particularly in a dialogical setting.

This study has made it clear that there are many subtle clues of stance taking in the dialogical interaction between speaking partners. Studies to date, this one included, focus

primarily on clues available in the current speaking turn. While I was able to incorporate clues from the previous turn, it is possible that there are even better contextual clues available within a wider window.

BIBLIOGRAPHY

Anand, P., Walker, M., Abbott, R., Tree, J. E. F., Bowmani, R., and Minor, M. (2011). Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis*, pages 1–9. Association for Computational Linguistics.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *LREC*, volume 10, pages 2200–2204.

Bell, A. (2009). Language style as audience design. In Coupland, N. and Jaworski, A., editors, *The new sociolinguistics reader*, pages 240 – 250. Palgrave Macmillan.

Biber, D. and Finegan, E. (1988). Adverbial stance types in english. *Discourse processes*, 11(1):1–34.

Biber, D. and Finegan, E. (1989). Styles of stance in english: Lexical and grammatical marking of evidentiality and affect. *Text-interdisciplinary journal for the study of discourse*, 9(1):93–124.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). Grammar of spoken and written English. *Harlow: Longman*.

Boersma, P. P. G. et al. (2002). Praat, a system for doing phonetics by computer. *Glot international*, 5.

Bradac, J. J., Bowers, J. W., and Courtright, J. A. (1979). Three language variables in communication research: Intensity, immediacy, and diversity. *Human Communication Research*, 5(3):257–269.

Bradley, M. M. and Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology, University of Florida.

Branigan, H. P., Pickering, M. J., and Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, 75(2):B13–B25.

Brennan, S. E. (1996). Lexical entrainment in spontaneous dialog. *Proceedings of ISSD*, 96:41–44.

Brennan, S. E. and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482.

Bruce, R. F. and Wiebe, J. M. (1999). Recognizing subjectivity: a case study in manual tagging. *Natural Language Engineering*, 5(2):187–205.

Brysbaert, M. and New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, 41(4):977–990.

Burger, S., MacLaren, V., and Yu, H. (2002). The ISL meeting corpus: The impact of meeting type on speech style. In *Seventh International Conference on Spoken Language Processing*.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., et al. (2005). The AMI meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

Cavazza, N. and Guidetti, M. (2014). Swearing in political discourse: why vulgarity works. *Journal of Language and Social Psychology*, 33(5):537–547.

Chafe, W. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. University of Chicago Press.

Choi, Y. and Wiebe, J. (2014). +/-EffectWordNet: Sense-level Lexicon Acquisition for Opinion Inference. In *EMNLP*, pages 1181–1191.

Clark, H. H. and Brennan, S. E. (1991). Grounding in communication.

Clark, H. H. and Schaefer, E. F. (1989). Contributing to discourse. *Cognitive science*, 13(2):259–294.

Clark, H. H. and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22(1):1–39.

Cleveland, W. S. (1981). LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *American Statistician*, 35(1):54.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., and Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708. ACM.

DeFrank, M. and Kahlbaugh, P. (2019). Language choice matters: When profanity affects how people are judged. *Journal of Language and Social Psychology*, 38(1):126–141.

Du Bois, J. W. (2007). The stance triangle. *Stancetaking in discourse: Subjectivity, evaluation, interaction*, 164:139–182.

Du Bois, J. W. (2014). Towards a dialogic syntax. *Cognitive Linguistics*, 25(3):359–410.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.

Freeman, V. (2015). *The phonetics of stance-taking*. PhD thesis, University of Washington.

Freeman, V. (2019). Prosodic features of stances in conversation. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 10(1).

Freeman, V., Chan, J., Levow, G.-A., Wright, R., Ostendorf, M., Zayats, V., Luan, Y., Morrison, H., Fox, L., Antoniak, M., et al. (2014). ATAROS technical report 1: Corpus collection and initial task validation. *U. Washington Linguistic Phonetics Lab.*

Freeman, V., Wright, R., and Levow, G.-A. (2015). The prosody of negative 'yeah'. In *LSA Annual Meeting Extended Abstracts*, volume 6, pages 6–1.

Garrod, S. and Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218.

Giles, H. and Powesland, P. (2009). Accommodation theory. In Coupland, N. and Jaworski, A., editors, *The new sociolinguistics reader*, pages 232 – 239. Palgrave Macmillan.

Gravano, A. and Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634.

Haddington, P. (2004). Stance taking in news interviews. *SKY Journal of Linguistics*, 17:101–142.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., and Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological science*, 22(1):39–44.

Jacobi, L. L. (2014). Perceptions of profanity: How race, gender, and expletive choice affect perceived offensiveness. *North American Journal of Psychology*, 16(2).

Jindal, N. and Liu, B. (2006). Identifying comparative sentences in text documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 244–251. ACM.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In *International conference on machine learning*, pages 957–966.

Lakoff, R. (1973). The logic of politeness; or minding your ps and qs papers from the 9th regional meeting of the chicago linguistic society. *Chicago: Chicago Linguistic Society*, pages 292–305.

Lakoff, R. T. (1989). The limits of politeness: theraputic and courtroom discourse. *Multilingua*, 8(2/3):101 – 129.

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 *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 236–241. IEEE.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:627–666.

Loper, E. and Bird, S. (2004). NLTK: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics.

Louviere, J. J. and Woodworth, G. G. (1991). Best-worst scaling: A model for the largest difference judgments. *University of Alberta: Working Paper*.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The Penn Treebank.

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., et al. (2005). The AMI meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88.

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184.

Mohammad, S. M. (2012). # emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 246–255. Association for Computational Linguistics.

Mohammad, S. M. (2017). Word affect intensities. *arXiv preprint arXiv:1704.08798*.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Nenkova, A., Gravano, A., and Hirschberg, J. (2008). High frequency word entrainment in spoken dialogue. In *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers*, pages 169–172. Association for Computational Linguistics.

Niederhoffer, K. G. and Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

Nir, B. and Zima, E. (2017). The power of engagement: Stance-taking, dialogic resonance and the construction of intersubjectivity. *Functions of language*, 24(1):3–15.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Pearson, K. F. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed

to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Pennebaker, J. W., Booth, R. J., and Francis, M. E. (2007). Linguistic inquiry and word count: LIWC [computer software]. *Austin, TX: liwc. net*.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Technical report.

Pinker, S. (2007). What the f***?

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). A comprehensive grammar of the English language.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raaijmakers, S., Truong, K. P., and Wilson, T. (2008). Multimodal subjectivity analysis of multiparty conversation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 466–474.

Rassin, E. and Heijden, S. V. D. (2005). Appearing credible? swearing helps! *Psychology, Crime & Law*, 11(2):177–182.

Řehůřek, R. and Sojka, P. (2011). Gensim—statistical semantics in python. *Retrieved from genism. org*.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., and Huang, R. (2013). Sarcasm as contrast between a positive sentiment and negative situation. In *EMNLP*, volume 13, pages 704–714.

Riloff, E. and Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 105–112. Association for Computational Linguistics.

Rocklage, M. D. and Fazio, R. H. (2015). The Evaluative Lexicon: Adjective use as a means of assessing and distinguishing attitude valence, extremity, and emotionality. *Journal of Experimental Social Psychology*, 56:214–227.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Sakita, T. I. (2006). Parallelism in conversation: Resonance, schematization, and extension from the perspective of dialogic syntax and cognitive linguistics. *Pragmatics & Cognition*, 14(3):467–500.

Scherer, C. R. and Sagarin, B. J. (2006). Indecent influence: The positive effects of obscenity on persuasion. *Social Influence*, 1(2):138–146.

Shriberg, E., Stolcke, A., and Baron, D. (2001). Can prosody aid the automatic processing of multi-party meetings? evidence from predicting punctuation, disfluencies, and overlapping speech. In *ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding*.

Slovenko, R. (1982). The impact of profanity on hearsay evidence. *Medicine & Law*, 1:397 – 402.

Somasundaran, S., Ruppenhofer, J., and Wiebe, J. (2007). Detecting arguing and sentiment in meetings. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, volume 6.

Somasundaran, S. and Wiebe, J. (2010). Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124. Association for Computational Linguistics.

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.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Thelwall, M. (2013). Heart and soul: Sentiment strength detection in the social web with sentistrength. *Proceedings of the CyberEmotions*, 5:1–14.

Thomas, M., Pang, B., and Lee, L. (2006). Get out the vote: Determining support or opposition from congressional floor-debate transcripts. *arXiv preprint cs/0607062*.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York, fourth edition. ISBN 0-387-95457-0.

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Wiebe, J. (2000). Learning subjective adjectives from corpora. *AAAI/IAAI*, 20(0):0.

Wiebe, J. and Mihalcea, R. (2006). Word sense and subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1065–1072. Association for Computational Linguistics.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning subjective language. *Computational linguistics*, 30(3):277–308.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210.

Wiebe, J. M., Bruce, R. F., and O'Hara, T. P. (1999). Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 246–253. Association for Computational Linguistics.

Wilson, T. (2008). Annotating subjective content in meetings. In *LREC*.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. (2005a). OpinionFinder: A system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations*, pages 34–35. Association for Computational Linguistics.

Wilson, T. and Wiebe, J. (2003). Annotating Opinions in the World Press. In *SIGDIAL Workshop*, pages 13–22.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005b). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.

Wilson, T., Wiebe, J., and Hwa, R. (2004). Just how mad are you? Finding strong and weak opinion clauses. In *aaai*, volume 4, pages 761–769.

Yu, H. and Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Linguistics.

Zeileis, A. and Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, 2(3):7–10.

Appendix A

TASK VOCABULARY

A.1 ATAROS 3I

boating supplies	fish hooks	heavy cable	tow rope
fishing nets	cook stoves	box knives	electric heaters
five-pound weights	siding	power cords	bundles of wood
half-inch tubing	axes	peat moss	mouse traps
bundles of sticks	matches	saw	duct tape
canvas bags	wet suits	cushions	hats
sweaters	coats	vests	boots
socks	jackets	toys	books
travel guides	flags	paper	scissors
chocolate bars	oatmeal	doughnuts	soy beans
beets	dried figs	shoelaces	pet food
refrigerator magnets	canned peas	egg timers	toilet paper
toothpaste	soap	face cream	tweezers
plastic jugs	buckets	paper bags	cups
bottled water	backpacks	sugar	bagels
cake mix	eggs	butter	cookies
ice cream	whiskey	juice	camping
fishing			

Table A.1: ATAROS 3I Task Vocabulary from Freeman (2015)

A.2 ATAROS 6B

towing services	speed limit signs	additional bus stops
taxi stops	boating licenses	junior soccer league
fishing licenses	bookkeeping classes	boys basketball club
football stadium upkeep	cooking classes	hunting tags
football equipment	reproductive education	reusable bag campaign
hospital additions	chicken pox vaccinations	STD education
needle exchange	veterinary hospital	egg farm regulations
pothole maintainance	weed control	subway system
invasive species removal	flag pole repair	public bus upkeep
drainage ditches	job training programs	teaching certificates
acting coaches	massage certificates	math tutors
tattoo artist licenses	sex offender database	stray cat spaying
toxic waste disposal	bagel factory inspections	kitten and puppy adoptions
dog catcher	pest control	soup kitchens
community news ads	prenatal checkups	housing assistance
veterans' medical assistance	food bank	public access station
neighborhood watch support	poetry books	sex ed
custodians	speech therapy	assistant cooks
special ed teachers	sugar-free juice machines	note-takers
disability services	music teachers	

Table A.2: ATAROS 6B Task Vocabulary from Freeman (2015)