APPROVAL SHEET

Title of Thesis: Identifying and Ordering Scalar Adjectives Using Lexical Substitution

Name of Candidate: Bryan Wilkinson

Doctor of Philosophy, 2017

Thesis and Abstract Approved: _____

James Oates

Professor

Department of Computer Science

and

Electrical Engineering

Date Approved: _____

ABSTRACT

Title of dissertation: IDENTIFYING AND ORDERING SCALAR ADJECTIVES USING LEXICAL SUBSTITUTION

Bryan Wilkinson, Doctor of Philosophy, 2017

Dissertation directed by: Professor James Oates
Department of Computer Science and Electrical Engineering

Lexical semantics provides many important resources in natural language processing, despite the recent preferences for distributional methods. In this dissertation we investigate an under-represented lexical relationship, that of scalarity. We define scalarity as it relates to adjectives and introduce novel methods to identify words belonging to a particular scale and to order those words once they are found. This information has important uses in both traditional linguistics as well as natural language processing. We focus on solving both these problems using lexical substitution, a technique that allows us to determine the best substitute word for a given word in a sentence. We also produce two new datasets: a gold standard of scalar adjectives for use in the development and evaluation of methods like the ones introduced here, and a test set of indirect question-answer pairs, one possible application of scalar adjectives.

IDENTIFYING AND ORDERING SCALAR ADJECTIVES USING LEXICAL SUBSTITUTION

by

Bryan Wilkinson

Thesis submitted to submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County in partial fulfillment
of the requirements for the degree of
DOCTOR OF PHILOSOPHY
2017

Advisory Committee:
Professor James Oats, Chair/Advisor
Professor Tim Finin
Professor Charles Nicholas
Professor Shimei Pan
Professor Mona Diab, George Washington University

To my parents, Lauren and Matt:

For a lifetime of support, encouragement, sacrifice, and love,

without which none of my achievements would be possible

and

To the members of the UMBC Community, past and present:

Who together have created an environment for myself

and thousands of other students to thrive

Acknowledgments

This PhD dissertation is not only a reflection of my work, but of all the people who have affected my life in one way or another leading up to this point. While I cannot possibly name everyone who has supported me along this journey, I wish to express my utmost gratitude to them all. I would like to take this chance to highlight several individuals and organizations I am especially thankful for.

First, I would like to thank my advisor, Dr. Tim Oates, who has supported me not only in my research goals, but also as a mentor during my first teaching experiences. Tim's support and generosity has allowed me to focus on my research. I am forever grateful to the example he has set as an advisor.

I would like to thank Drs. Sergei Nirenburg, Marjorie McShane, and Stephen Beale, whose knowledge and experience I benefited from greatly as both an undergraduate researcher and as a a new graduate student. I also wish to thank Drs. Jesse English and Ben Johnson, for modeling what graduate work should be like. Additionally I would like to thank Drs. Tim Finin, Charles Nicholas, Shimei Pan, and Mona Diab for serving on my committee and their invaluable feedback on this work.

I would like to thank Dr. Milt Halem and his students in the CHMPR lab for setting me up with account on their MINSKY machines. Access to this resource allowed this dissertation to proceed at a much more efficient pace.

I must also thank Professor Alan Bell of the UMBC Modern Languages and Linguistics department, with whom I met with at the behest of my grandfather,

Jack Cohen. Dr. Bell introduced me to the discipline of Applied Linguistics, and suggested I minor in it as an undergraduate student, setting off a series of events that led to my research in Natural Language Processing.

Throughout my life I have been lucky to have many wonderful teachers. I want to specifically thank Mrs Joyce Crompton, for teaching me to love learning and school; Mrs. Trish Irwin for instilling in me a love of science and discovery; Mrs. Linda Leonard for showing me the power that a friendly face can have on a young student; Mrs. Amy Sainz for introducing me to Spanish and my love of languages; and Mr. Tom Marvel and Mrs. Patricia Villani, for always making me feel welcome and safe at school.

I would like to thank the staff of the Damascus Library, in Damascus, Maryland, for always helping me find something new to learn about, and fostering my love of reading.

I would like to thank two scout leaders from my youth, Ron Bridge and Doug Coe, for showing me what it means to be a good person and treating me like another son. I would also thank their families, Janet, Joseph, Jonathan, and Joshua, as well as Marian, Jennifer, and David for the many years of support, friendship, and interest in my work.

I wish to thank Drs. Brandon Borde, Ejiofor Ezekwe, and Steve Tuyishime, as well as soon-to-be-doctor Ramon Cabrera, whose tireless work and love of research inspired me while we were undergraduates together. I could not have fallen in with a better group of people my first week at UMBC.

While a graduate student at UMBC, I was fortunate enough to remain involved

in campus life through the Student Events Board (seb). While listing every person I met through this group would take an entire page on its own, I want to thank each of you for your influence. I want to particularly thank Ms. Jen Dress and Dr. Lee Hawthorne, who allowed me to remain connected to the UMBC community as a graduate student.

I also owe thanks to Mrs. Courtney Haupt, who was kind enough to read this dissertation and provide grammatical edits. In addition, I would like to thank my fellow students in the CoRAL lab, who have listened, discussed ideas, and offered suggestions numerous times over the years: Drs. Neils Kasch, Huegens Jean, and Zhiguang Wang, as well as John Clemens, Sunil Gandhi, Ashwin Ganesan Hang Gao, Kavita Krishnaswamy, Akshay Peshave, Craig Pfeifer, and Chi Zhang.

Finally, I wish to thank my family. I am so lucky to have you all around me. I wish to thank my late great-grandmother Dorothy Stein, and her siblings, for teaching me the importance of remaining close as a family, even as you age. I also wish to thank my late grandfather, Tom Wilkinson, who taught my how to use a microfiche reader and the joys of searching for answers where ever they may be found.

Thanks to my extended family: Michael, Nadine, Dylan, Jan, Adam, Surena, Kira, Tillula, and Jacinda. You have loved and supported me during this dissertation as only family can.

Thanks to my grandparents, Ina and Jack Cohen, who have been a constant presence in my life, and have shown me the importance of doing what you love as a career. I am thankful to have had you so close during my time at UMBC.

Thanks to my siblings, Tyler, Jordann, and Avery. I am truly lucky to have grown up with you, and know how special our bond is. Your encouragement, understanding, and love has meant the world to me.

Most importantly, I owe the deepest gratitude to my parents, Lauren and Matt, who have been the greatest role models I could have possibly wished for in life. Mom, thank you for instilling in me a love of teaching. And Dad, thank you for instilling in me a love of computers. Besides your direct influences on my studies, I could not have made it through the demanding years as a graduate student without your love, care, and advice.

Table of Contents

List of Tables x

List of Figures xii

1 Introduction 1
- 1.1 Contribution . . . 3
- 1.2 Significance . . . 3
- 1.3 Outline . . . 4

List of Abbreviations 1

2 Background 5
- 2.1 Adjective Meaning in General . . . 5
- 2.2 Antonyms and Their Relation to Scales . . . 8
- 2.3 Scales . . . 10
  - 2.3.1 Scale Membership . . . 10
  - 2.3.2 Scale Structure . . . 13
  - 2.3.3 Attested Scales . . . 17
- 2.4 The Linguistic Importance of Scalar Adjectives . . . 19
- 2.5 Computational Semantics of Adjectives . . . 20
  - 2.5.1 Distributional Semantics . . . 20
  - 2.5.2 Knowledge Based Semantics . . . 23
- 2.6 Previous Approaches to Constructing Scales . . . 25
  - 2.6.1 Scale Membership . . . 25
  - 2.6.2 Scale Ordering . . . 27
- 2.7 Lexical Substitution . . . 35
- 2.8 Summary . . . 37

3 Data 38
- 3.1 Introduction . . . 38
- 3.2 Gold Standard . . . 38
  - 3.2.1 Previous Datasets of Scalar Adjectives . . . 40
  - 3.2.2 A Gold Standard for Scale Membership . . . 41

3.2.2.1 Methodology . . . 42
3.2.2.2 Results . . . 44
3.2.2.3 Toy Example . . . 48
3.2.3 A Gold Standard for Scale Ordering . . . 51
3.2.3.1 Results . . . 52
3.2.4 Comparison against other hand created data sets . . . 54
3.2.5 Discussion . . . 55
3.2.6 Future work . . . 56
3.2.7 Conclusion . . . 57
3.3 More Indirect Question Answer Pairs . . . 57
3.3.1 Constructing a Diverse IQAP . . . 60
3.3.2 Results . . . 62
3.4 Summary . . . 65

4 Identifying Scale Members 66
4.1 Methodology . . . 66
4.1.1 Pattern-Based Exemplars . . . 67
4.1.2 Lexical Substitution . . . 69
4.1.3 Cultural Consensus Theory . . . 70
4.2 Experimental Setup . . . 71
4.3 Hpyerparameter Anyalsis . . . 72
4.4 Evaluation . . . 76
4.5 Discussion . . . 79
4.5.1 Expansion . . . 80
4.5.2 Length . . . 82
4.5.3 Trial . . . 84
4.6 Summary . . . 85

5 Ordering Scale Members 86
5.1 Methodology . . . 87
5.1.1 Partitioning the Scale . . . 87
5.1.2 Augmentation . . . 88
5.1.2.1 Normalization . . . 89
5.1.2.2 Independent or Combined Normalization . . . 90
5.1.2.3 Sampling Proportionally . . . 90
5.1.3 Ordering . . . 91
5.1.4 New Model . . . 93
5.2 Experimental Design . . . 94
5.3 Results . . . 95
5.4 Application to Data from (de Melo and Bansal 2013) . . . 103
5.5 Summary . . . 105

6 Learning and Ordering Scales from Noisy Seeds 106
6.1 Determining Scale Membership . . . 106
6.1.1 Selecting Seed Words . . . 106
6.1.2 Pruning Scales . . . 108
6.2 Human Evaluation of Membership . . . 108
6.2.1 Scales from WordNet seeds . . . 109
6.2.2 Scales from VM Seeds . . . 111
6.3 Comparison of Membership with Existing Resources . . . 113
6.4 Determining Scale Order . . . 114
6.5 Human Evaluation of Scale Order . . . 115
6.6 Indirect Question Answer Pairs . . . 117
6.6.1 Prior Work . . . 118
6.7 Summary . . . 119

7 Conclusion 120
7.1 Limitations . . . 120
7.2 Future Work . . . 121

List of Tables

2.1 Patterns used in (van Miltenburg, 2015) . . . 26
2.2 Patterns used in (Sheinman and Tokunaga, 2009) . . . 28
2.3 Patterns used in (de Melo and Bansal, 2013) . . . 28

3.1 Eigenvalue ratios for 16 sets of words, proposed scales above the line and sets of adjectives that do not make up scales and were used as controls below the line. . . . 47
3.2 G_k for words along two postulated scales (a, b) and one set of adjectives that describe material (c). Words marked with $*$ were prompt words. 48
3.3 Example responses for toy example. . . . 49
3.4 Scale orderings and the corresponding eigenvalue ratios . . . 53
3.5 τ_k for words along the scales for SIZE (a), TEMPERATURE (b), and INTELLIGENCE (c) . . . 54
3.6 Most frequent adjectives found in questions in IQAP dataset . . . 59
3.7 Most frequent adjectives combinations found in IQAP dataset . . . 59
3.8 Examples question-answer pairs demonstrating pruning process. Pairs with PMI less than 3 are pruned. . . . 64

4.1 Dependencies of the original and substituted words in sentence 4.15. The dependencies appear much more plausible than the sentence as a whole. $^{-1}$ represents an inverse dependency, that is the second word listed is dependent on the word being discussed . . . 68
4.2 First 5 suggested substitutes for sentences (4.15) - (4.17) using the three algorithms studied. . . . 70
4.3 ANOVA and η_p^2 with F-score as the dependent variable. Lines in bold represent significant parameters . . . 73
4.4 Results of our **P0.1N** and **No10C** configurations compared to (van Miltenburg, 2015) and a simple word embedding model on the development set . . . 78
4.5 Results of our **P0.1N** configuration compared to (van Miltenburg, 2015) on the test set . . . 78
4.6 Frequency of seed pairs found in batches with *bad,awesome* . . . 85

5.1 Patterns used to orient the two subscales relative to each other. x is the word in the right subscale. . . . 91
5.2 Results of ANOVA using ρ as the dependent variable. Bold lines are statistically significant . . . 96
5.3 Performance in terms of ρ of ordering methods . . . 97
5.4 Performance in terms of Pairwise Accuracy of ordering methods . . . 98
5.5 Percent of new sentences generated that demonstrate the correct relationship . . . 98
5.6 Best Configuration given Scale Structure . . . 101
5.7 Best ρ given scale structure . . . 101
5.8 Pairwise accuracy given scale structure. Represents value cannot be determined because linear solver cannot be run . . . 102
5.9 Difference in correctness, density, and percentage of new sentences between lexical substitution and random substitution for each type of scale structure. . . . 102
5.10 Difference in correctness, density, and percentage of new sentences between lexical substitution and word2vec context-free substitution for each type of scale structure. . . . 102
5.11 Performance on dataset from (de Melo & Bansal 2013) . . . 104

6.1 Sample scale memberships using two different seed sets. Bold words are seed words while italicized words were indicated as incorrect by a human annotator . . . 112
6.2 Percentage of new words discovered using membership identification methodology . . . 113
6.3 Average number of attributes and frames per scale. . . . 114
6.4 ρ for human evaluation of scales. . . . 116
6.5 Accuracy on IQAP and MIQAP testsets. . . . 118

List of Figures

1.1 The scale for SIZE as determined by this dissertation 2

2.1 Wordnet structure for adjectives of QUALITY 24

3.1 Overview of methodology. 40
3.2 Lexical elicitation interface. 44
3.3 Comparing informants with different seed words 46
3.4 Item-by-informant matrix. 49
3.5 Informant-by-informant correlation matrix. 49
3.6 Competency Vector. 50
3.7 Adjective Ordering Interface. 51
3.8 Interface used to gather indirect responses. 61

6.1 Interface used by informants to indicate which words do not belong to each scale. 109
6.2 Interface used by informants to indicate which scales are ordered correct. 116

Chapter 1: Introduction

Lexical Semantics is the study of the meaning of words and the relationships between those meanings. This is usually encoded for computational use in a graph format, like WordNet (Fellbaum, 1998), or in a frame structure, like FrameNet (Baker et al., 1998). Recently the increased use of distributional semantics has shifted lexical semantics to the practice of predicting relations between two words, rather than hardcoding them in some knowledge base.

Even with the recent success of distributional semantics, traditional lexical semantics in the form of encoded knowledge is still important and useful. Distributional semantics have been enhanced by retrofitting vectors using knowledge sources, and it has been shown that even a semantic network itself can be encoded into a vector (Faruqui et al., 2015; Faruqui and Dyer, 2015).

Lexical semantics depends on the relationships between the words, and these relationships have traditionally focused primarily on nouns, somewhat on verbs, and very little on adjectives. Consider the relationships in WordNet, hypernonmy and hyponymy, which encode the "IS-A" relationship, usually between two nouns, although sometimes between two verbs as well. This makes up 72.66% of the relationships in WordNet.

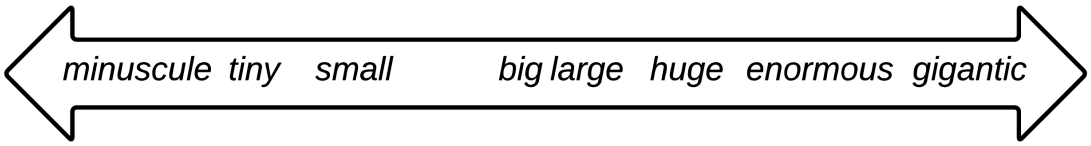

Figure 1.1: The scale for SIZE as determined by this dissertation

In this dissertation, we will look at a long attested relationship, grading, which occurs when two adjectives occur on the same scale, which we define as adjectives representing a property to different degrees, and one adjective is more intense than the other (Sapir, 1944; Casagrande and Hale, 1967; Cruse, 2011). Figure 1.1 shows the scale for SIZE, including the words that modify size as well as their relative ordering. Previous work on this issue has focused primarily on ordering the words on a scale using corpus data (Sheinman et al., 2013; De Melo and Bansal, 2013; Ruppenhofer et al., 2014). These works take the words on the scale as input, using the WordNet similar-to relationship or FrameNet frames as the set of words to order.

While there has been limited work on determining which words are on a scale, many approaches use clustering, which requires that the number of scales is known ahead of time and that all the words to be clustered belong on a scale (Hatzivassiloglou and McKeown, 1993; Shivade et al., 2015).

Apart from this dissertation, the only works known to perform both these tasks are van Miltenberg (2015) and Shivade, et al. (2015). Van Miltenberg uses pattern based methods to find pairs of words, and then orders these pairs of words. Shivade, et al. report the results of running the MILP method of (De Melo and Bansal, 2013) on their clusters. In this work we produce a system which given two words, provided

by a user or the output of a method like (van Miltenburg, 2015), gathers the words that are on the corresponding scale and then orders them.

1.1 Contribution

As a result of this work, we produce ScaleBundle, a resource of the learned scales, freely open to the public for use. Additionally, we present two datasets, one of which is a gold standard of the membership and ordering of 12 scales. The other dataset presented is an extension of the Indirect Question Answer Pair (IQAP) dataset.

While the primary contribution of this work is a system to identify and order adjectives, in doing so we have demonstrated several novel uses of existing techniques. We use the lexical substitution task, traditionally seen as a way to evaluate context sensitive word embeddings, as a way to augment explicit patterns found in text. We also use the lexical substitution task to estimate what words could be found in a pattern, but aren't.

In addition, we demonstrate that cultural consensus theory, a technique from test theory and anthropology, can be used to generate crowd-sourced labels with very limited prior knowledge of correct labels.

1.2 Significance

The primary significance of this work is achieving state of the art performance on both the identification of scale members as well as their ordering. As a result of

this, we are able to achieve improved performance on the indirect question answering task. Finally, we produce a large resource of scales that will enable linguists and others to further investigate the theory behind scalar adjectives.

1.3 Outline

Chapter 2 of the dissertation covers the background on prior work done on adjectival scales as well as the linguistic motivation for their existence. We will also cover the basics of distributional and pattern based semantics in this chapter. Chapter 3 discusses the construction of both of our datasets, as well as the details of Cultural Consensus Theory. Part of this work was presented at LREC 2016, as (Wilkinson and Oates, 2016). In Chapter 4 we present our system for determining scale membership, and evaluate it on the dataset described in chapter 3. Chapter 5 gives the details of our system to order adjectives. This system takes in any list of words, and assumes that they are on a scale. We evaluate the ordering system in isolation in this chapter, but in Chapter 6 we explore an end-to-end system using both methodologies along with noisier input. Also in Chapter 6, we present one potential use case of this new resource, an improved system to solve the IQAP dataset. In Chapter 7 we present a conclusion and potential future research directions.

Chapter 2: Background

In this chapter we present current linguistic theories on adjectives as a class of words and then discuss several subclasses of adjectives. Having done this, we define the specific set of adjectives our work will focus on, and review relevant theory about adjective scales and scalar adjectives. We next review computational methods of modeling meaning as well as related work to the dissertation. Finally, we review relevant literature specific to our solution to the problem.

2.1 Adjective Meaning in General

Adjectives have been studied since the earliest linguists, although when they were recognized as a word class is less clear, having been lumped in with nouns for most of recorded history. It appears that by the 13th century a specific subclass of nouns was at least recognized by Western European linguists (Householder, 1995; Bursill-Hall, 1995). The Persian linguist, Sībawayhi, identified a class of words in Arabic as "describing words" even earlier, by the 8th century (Carter, 1973; Thomas, 2011).

Unlike the more studied word classes, nouns and verbs, the universality of adjectives is not agreed upon (Sasse, 1993). Dixon proposes that all languages have

the adjective class but the size of the class may be rather small in some languages, to the point where adjectives could be considered a closed class in that language, like prepositions are in English (Dixon and Aikhenvald, 2006). Cruse contends properties of concepts may be modified by other parts of speech in some languages and that the adjective class is not universal (Cruse, 2011). This is important to note as it implies that the techniques presented in this dissertation, while theoretically language independent, may need specific modification to work with languages other than English.

The exact definition of an adjective is not clear even in a language such as English which is widely accepted to have them. Trask's semantically oriented definition that "the meaning of an adjective is most typically a temporary or permanent state..." highlights the imprecise nature of the literature (Trask, 1998).

To orient the reader with the subset of adjectives we will work with for the remainder of this dissertation, we will now briefly review subtypes of adjectives before homing in on scalar adjectives.

In English, adjectives can be described syntactically by the positions they occupy relative to the noun they modify: predicative, attributive, or both. Predicative adjectives appear as the object of a verb such as *is* while attributive adjectives appear directly before the noun they modify (Kennedy, 2012). Prior linguistic research may restrict their studies to one or the other, but we make no such distinction.

Semantically, classification with regards to the effect on compositional meaning with the noun modified is common. One such division is intersective and subjective, two highly intertwined categories of adjectives. Intersective adjectives have meanings

which reflect the intersection of the domains of the adjective and the noun while subjective adjectives create a subset of the noun. A canonical example of an intersective adjective is *red*, while perhaps the most cited example of a subjective adjective is *beautiful* (Abdullah and Frost, 2005). Abdullah notes the categories are not always distinguishable while Kennedy goes as far as to place intersective adjectives as a subset of subjective adjectives. These can be contrasted with non-subjective adjectives, which often negate a property, like *fake* or *former* (Kennedy, 2012).

Critically, intersective and subjective adjectives can be additionally described as gradable or non-gradable. Gradable adjectives have meanings related to some scalar property and often relate to a standard, although as discussed in section 2.3 this is not universally accepted.

Paradis further breaks down the class of gradable adjectives into scalar and non-scalar adjectives. Non-scalar adjectives do not have antonyms but have what Cruse terms complimentaries, such as *dead* and *alive*. These adjectives are still gradable because they combine with a limited set of adverbs which Paradis calls totality modifiers of degree, such as *totally*. Adjectives that do occur on scales are broken down into two groups: those that represent ranges on the mental scale, which Paradis also terms scalar, and those that represent a single, and more extreme, point on the mental scale, such as *excellent*, which Paradis calls extreme adjectives (Paradis, 2001). For the purposes of this dissertation, we will use Paradis' more permissive definition of a scalar adjective that includes both extreme adjectives and the traditional scalar adjectives like *good* and *bad*.

2.2 Antonyms and Their Relation to Scales

Before discussing theories of scales, we give a brief overview of theories of antonymy, which while certainly intertwined with scalararity is by no means clearly a part of it (Bolinger, 1977).

Gross, et al. investigated the mental representation of adjectives (Gross et al., 1989). From a pyscholinguistic perspective, antonymy is considered by some linguists as central to the organization of adjectives within the mind, and thus their meaning. The work of Gross, et al. is limited to adjectives in the predicative position only. They concluded both antonymy and synonymy are required for organization of adjectives. This model of adjective meaning has been implemented in WordNet and is commonly referred to as the dumbbell model, which we discuss in more detail in section 2.5.2.

Ljung provides an excellent overview of different definitions of antonymy in the literature (Ljung, 1974). Zimmer and Lyons give the description most relevant to this discussion, that antonymous adjectives are opposites "along some given dimension" (Ljung, 1974). Ljung also concludes that an antonym pair's validity depends on a particular person's experience and interpretation of the world. As an example, not all people may consider *beautiful* and *ugly* as antonyms. This should not preclude the words from being on the same scale. From this purely theoretical perspective, we can see that adjectives are considered antonyms in this model if they are opposites on the same scale.

Leher and Leher also comment on scales' relationships with antonyms. One way

to look at antonymy is antonyms must occur on the same scale and be equidistant from the midpoint of the scale. They argue that antonymy is a function of the existence of scales rather than an integral part in forming them, a viewpoint we adopt (Lehrer and Lehrer, 1982).

Taking a corpus based approach, Muehleisen investigates why words which she terms "near-opposites" are not considered antonyms (Muehleisen, 1997). In a large corpus constructed from New York Times articles and Project Gutenberg files, it was found that the collocation patterns with nouns must be common for two words to be considered antonyms. Thus, *large* and *little* are not considered antonyms as they do not collocate with many of the same nouns. This finding suggests that basic frequency counts are not enough in themselves to determine scales.

Raybeck and Herrman investigate antonymy across 10 different languages. Participants were presented three pairs of words that the authors believed to represent a semantic relationship, such as contradictory opposites or synonymy. They were asked to group these relationships together. From this they determine that the different types of opposites are grouped together most consistently (Raybeck and Herrmann, 1996). While they are hesitant to pronounce antonymy as universal, they say their results are highly suggestive. This result only holds weight, however, if we accept that antonymy is defined as three different types of oppositness: contradictory, reverse, and directional. If we do, we can combine this result with the assumption of Lehrer and Lehrer's that antonyms rely on scales to conclude that scalarity must also be universal.

2.3 Scales

We begin this section by going over the linguistic literature concerning scales and then conclude with a review of work on specific scales.

Work on lexical scales can be traced back to the work of the seminal linguist Edward Sapir. Sapir notes that some sets of words can be ordered into a sequence. The basis for these orderings can be done on three levels: logical, psychological, or linguistically. His large typology of these orderings shows that producing ordered series of words is not an easy task (Sapir, 1944).

Work subsequent to Sapir can be divided into discussions of scale membership and scale structure. While some literature addresses both, we find it helpful to introduce the relevant literature thematically, addressing membership first.

2.3.1 Scale Membership

Bolinger begins his discussion of what he terms degree words, scalar adjectives in our terminology, by noting that in antonymous pairs one is neutral and the other is biased. The nuetral member of the pair is capable of referring to any part of the scale in the abstract, while the biased word only refers to a particular value of a property. For example, in the pair *old-young*, *old* is the neutral member, while *young* is the biased word. Similarly, in the pair *strong-weak*, *strong* is neutral and the biased member is *weak*. This is similar to the concept of having a marked and unmarked member of the pair (Bolinger, 1977). The neutral member of the pair is used when asking questions and in reference to the property in general. He goes on to say that

this notion can be extended to "multi-term relationships" (Bolinger, 1977) with one neutral member and several biased members. Boligner makes extensive use of "*how* questions" which, when one word answers are given, can provide interesting insight into their scalarity. When the how question contains a scalar adjective, Bolinger says that this restricts a one word answer to a degree of that scale. An example given is the question "How good is George as an accountant?" He claims *bad* and *good* are not possible answers to this question. However possible answers include: *ok*, *fine*, *great*, and *terrible*. Bolinger does not address the fact that *bad* is the antonym for *good* but is not biased as it cannot answer the how question. He does posit that the allowed responses are all degree words of the adjective in question, while *good* and *bad* are not degree words themselves. It is not clear if the example questions given were constructed for the purposes of his argument or were found in an empirical study.

Bolinger artfully captures the difficulty of defining what a scale is as follows:

> The loosely related sets include pairs whose members sometimes seem as if they had met by accident; there are better and worse choices for an antonym, and speakers may disagree and writers lucubrate – *polite*, for example, as *impolite* as its mate, and most would agree on *rude* as a companion, but leftward of *rude* one might or might not put such terms as *boorish*, *churlish*, and *uncouth*; and it is not easy to decide how to orient *courteous-discourteous* and *mannerly-unmannerly* to the *polite-impolite* scale. The adjectives and adverbs of English are an unruly tribe, with a

few established marriages among a host of summer romances.

> Some of the linear sets are complex, with members spaced out towards both extremes of the scale and serving as intensifiers of the members farther in. But other scales exist, and it is not even necessary that there be a typical antonymous pairing (Bolinger, 1977).

Continuing with the examples given by Bolinger, in this dissertation we seek to answer using computational linguistics if *boorish* and *churlish* belong on the scale containing *polite* and *rude*; we seek to place *courteous*, *discourteous*, *mannerly*, and *unmannerly* in the correction position on the *polite-impolite* scale.

Westney attempts to enumerate the exact criteria that must be met for a word to be on a scale, listing and thoroughly discussing his suggested criteria for membership on scales. He concludes that gradability cannot be required of all members of a scale, citing the example scale ⟨*necessary*,*probably*,*possible*⟩ and noting that *possible* is not a gradable adjective.

The most important property for members of a scale to have according to Westney is incompatibility; they must be similar enough to be able to form relations with the other members of the scale, but still contrast so that they are not redundant. After giving numerous suggestions for criteria to determine scale membership, he concedes that "there is no obvious sufficient condition which would determine what scales are and what their membership is" (Westney, 1986). From this Westney concludes there is no claim to scalarity as a set of lexical items, as it would be "an endless task" to enumerate all the members. While this may be true, it should not

preclude us from attempting to define the parts of scales we can. Speakers of a language intuitively feel that groups of words are related, and knowledge of this is used in many situations in daily conversation, such as scalar implicature.

Taking a formal semantics approach, Kennedy argues that gradable adjectives are functions from objects to intervals on a scale. Each adjective operates on a domain of objects, and imparts an ordering among them, e.g., ordering a set of books based on their length, given that Les Miserables is described as *a long book* and The Hound of the Baskervilles is described as a *short novel*, the order would be ⟨ The Hound of the Baskervilles, Les Miserables⟩. He provides an excellent analysis of how an individual word's meaning may be interpreted, but little analysis of an entire scale as a unit. He touches on antonymy, but again from the standpoint of groups of objects having a particular ordering, not how words might relate to each other on the scale (Kennedy, 1997).

2.3.2 Scale Structure

Turning our attention to the structure of scales, in the same work previously discussed, Bolinger proposes many properties of scales. The first is that they are usually infinite. The evidence for this is that no matter the term, a speaker can always modify it with an adverb like *more*, thus moving the term's position on the scale further away from the center (Bolinger, 1977). This has been contested by Kennedy and McNally, who, while conceding that for many scales the end points are infinite, provide several examples of scales that are not (2005). They propose 4

possible classes of scale structure:

- Totally Open Scales - the structure Bolinger proposes, for scales like *short-tall* and *deep-shallow*.
- Lower Closed Scales - scales that have a definitive end point on the lower end of the scale, such as *unknown-famous* and *quiet-loud*.
- Upper Closed Scales -scales that have a definitive end on the high end of the scale, such as *dangerous-safe* and *uncertain-certain*.
- Totally Closed Scales - scales where both ends are definitive, such as *empty-full* and *closed-open*.

Bolinger himself seems to hint at these differences, noting that a scale for a particular color could be constructed, such as ⟨*pink, red, crimson*⟩, which is only infinite in one direction and contain no antonyms.

The difference between upper closed scales and lower closed scales requires that each scale have a positive and a negative end. Bolinger refers to this property as the orientation of the scale. Several different hypotheses as to the cause for this exist. One proposed explanation is that one direction of the scale is the norm for a property, such as *clean*. This explanation is based on the value society places on that property, and may be related to sentiment as used in NLP.

Another way to determine the positive direction is as a function of how easy it is to think of words for that dimension. To give a concrete example, Bolinger uses the scale that corresponds to *dim-bright*, stating it is much easier to think of words

closer to *bright*, such as *brilliant*, *vivid*, *light*. The side of the scale containning *bright* therefore is considered the positive dimension.

Related to this, Bolinger suggests that the frequency of a word as encountered during childhood influences which direction is seen as positive. His example of this is the scale *small-big*, which he claims is ordered this way because *big* is encountered by the child at a younger age and more often.

Aside from determining which end is the positive end, it is also important to determine where positive turns into negative on the scale, that is, what is the midpoint. Not all midpoints of a scale will be lexicalized, but when they are we can use the test proposed by Lehrer and Lehrer. They note that the midpoint of a scale cannot be modified by the adverbs *more* or *less*. The particular example they cite is *lukewarm*, especially as used when describing liquids (Lehrer and Lehrer, 1982).

The previous discussion has made the assumption that scalar adjectives occur on a single dimension. While this is a logical assumption to make, it is not a universally agreed upon truth.

Going beyond the scope of scalar adjectives, Cruse describes the properties that adjectives in general modify as "prototypically unidimensional, denoting an easily isolatable concept, such as length or temperature, in contrast with prototypical nouns, which denote rich highly interconnected complexes of properties" (Cruse, 2011). When studying a specific adjective it is easy to conceive of it as to one dimension, e.g., the dimension of HAPPINESS for the adjective *happy*.

Analyzing a group of related scalar adjectives together leads to a murkier picture. Lehrer and Lehrer contend that it is far more common to encounter

something more complex then a single dimension, listing sets such as {*happy*, *sad*, *angry*, *frustrated*} , {*clever* ,*smart, skillful, dumb, dull*}, and {*beautiful, pretty, ugly, homely*} as examples (Lehrer and Lehrer, 1982). {*happy*, *sad*, *angry*, *frustrated*} may be a set of incompatible emotional adjectives that should be treated similar to how Bolinger treats colors, while {*beautiful, pretty, ugly, homely*} present issues in their analysis because *homely* is much more distributionally restricted, occurring with fewer nouns. For the purposes of this dissertation, this distinction is less important as the interpretation of *homely* when combined with a noun still results in the noun's property of AESTHETICS being changed.

Somewhat between these two viewpoints, Beirwisch breaks down the group of gradable adjectives into dimensional adjectives and evaluative adjectives. These different classes of gradable adjectives have different scale structures. Dimensional adjectives, such as *long, short, old,* and *new*, with opposite meaning are still placed on the same scale, while evaluative adjectives, such as *pretty* and *ugly* are placed on completely separate scales. There are numerous formal semantic reasons for this, but perhaps the most intuitive one is that an object described as *short* still has height, while one described as *ugly* could be interpreted as having no beauty. Another difference between them is that only dimensional adjectives can be compared to the average for the comparison class they represent, while evaluatives do not have an average value to be compared to (Bierwisch, 1989).

2.3.3 Attested Scales

Surprisingly, for all this theory, very few scales have been thoroughly researched or even posited (Van Tiel et al., 2016). Some work in this area has been done by those interested in constructing marketing surveys and a limited amount has been carried out by linguists. One famous area of study has been color, but it is less clear if colors are scalar adjectives (Berlin and Kay, 1969). One promising area of work is the study of temperature lexicons.

Sutrop was among the first to look at temperature in detail, using the work of Berlin and Kay as a guide. Applying methods developed for the statistical elicitation of color terms, he preformed a study on 80 speakers of Estonian. The study consisted of four tasks, though only three are relevant to our work. First the participants were asked to name all the words for temperature that they knew. There was no mention of the participants being asked to only name adjectives, but their responses and the next task seem to indicate this was the result regardless. When analyzing the responses, several thresholds were used to determine a word's basicness. Sutrop adapts the concept of basicness from Berlin and Kay, defining it explicity for temperature terms as being "psychologically salient, [...] morpohlogically simple, [... and] applicable in animate, inanimate, and weather domains" (Sutrop, 1998).

The thresholds used were that over 50% of informants must list a word, that the word's position in a list has a mean of less than 4, and a threshold on the salience of a word, a computation proposed by Sutrop that combines the list frequency and average list position into a single score. All thresholds were determined based on

where the largest step occurred in several measurements. It was found that from the free listing exercise *külm(cold), soe(warm), kuum(hot), jahe(cool), palav(burning)* and *leige(tepid/lukewarm)* are potential basic words for temperature in Estonian (Sutrop, 1998).

In the second task, each participant was asked to name antonyms of each word they listed in the first task. Sturop contends on the basis of these results, only *külm, soe*, and *jahe* are potential basic words.

In the final task participants ranked the words from the list task and the antonym task by assigning each word either a negative or positive value. It is interesting that over 60 of the participants included *kuum* on their ranks yet it was given as an antonym so few times. This suggests that while antonymy is useful for elicitation of additional terms and for thinking about scales, it cannot be used definitively in determining scale membership.

Sturop also provides evidence against the idea that antonymy is conditioned on two words being equidistant from the center of the same scale. *Külm*'s average value was -2.44 while *kuum*'s value was 3.08. While interesting, this is not definitive proof, as each individual only ranked the words they produced in the earlier task. If the ranking task had been carried out on the same set of words for everyone, a more definitive picture could have been painted.

Koptjevsakaja-Tamm and Rakhilina investigate temperature words in Sweedish and Russian. Using the two languages for contrasts, they point out many different nuances a temperature system can have. For example, a different word is used based on whether the heat is sensed by touch or through the air in Russian. Both languages

have no basic words for extremely cold temperatures, although their requirements for basic words are not mentioned explicitly. The most interesting idea from this paper is that while a word can have a default antonym, the context may elicit a different antonym. The example presented is that the antonym for *cold* when used to describe beer is *lukewarm* rather than *hot* (Koptjevskaja-Tamm and Rakhilina, 2006).

The preceding overview of the linguistic investigations into this phenomenon should make clear that there is no consensus on this subject. While many of the prior studies look at antonymy and synonymy, what we are investigating is a possible set level relation.

2.4 The Linguistic Importance of Scalar Adjectives

In addition to the general desire to further understand the semantics of scalar adjectives, they have been studied with regards to at least two other linguistic phenomena, both of which have potential NLP applications.

The first area is what linguists term the study of binomials. A binomial is defined as two words of the same grammatical category joined with the word *and*, such as *salt and pepper* (Benor and Levy, 2006). Speakers of a language show a strong preference for the order of the two words, so much so that some phrases are considered frozen. The binomial construction is productive however, and all binomial constructions obey a set of constraints on their ordering. Among the many potential constraints investigated by Benor and Levy are POWER and SCALAR SEQUENCING. The power constraint requires the more powerful word to come first

in the construction, such as *horse and buggy*. It also can be interpreted as the more intense adjective should come first, as seen in *rich and poor*. In contrast, the scalar sequencing constraint requires that if two items exist in a sequence, they must appear in that order in the construction. These two constraints can be opposing, and it is an ongoing question as to why some binomials appear to obey one while others obey the other. An increased inventory of scales may advance work on this question.

The second and more widely known application of scalar adjectives is scalar implicature. This phenomenon is seen in a phrase like "some of the students passed the test." The hearer of this phrase knows that not all students passed the test, due to the Gricean maxims of quantity, which requires speakers to only say as much as necessary, and their knowledge of the scale, that is, what alternative adjectives could have been used, but were not.

2.5 Computational Semantics of Adjectives

How to represent the meaning of words, sentences, and documents has long been a goal of natural language processing. In this section we will review approaches to this, focusing on the meaning of words only. We split the review into two categories of representation, distributional based and knowledge based representations.

2.5.1 Distributional Semantics

Representing words with vectors has a long history, based on the notion that neighboring words provide enough information to estimate the meaning of the word

in question. This concept hypothesis is often attributed to Harris (1954) and Firth (1957). The practice of representing words this way is known as both distributional semantics as well as vector space models.

In the most basic vector space model, each word is represented with an $|V|$-length vector, where $|V|$ is the vocabulary size of the corpus. Each entry at index c in the vector representing the word w is the number of times the context word c appeared within a k length window around w in the corpus used to create the vectors. This results in very large and sparse vectors (Turney and Pantel, 2010).

One variation of this simple method is to use all dependents and heads of a word as determined by a dependency parser as the context for a word, rather than a fixed window size. This results in a word vector that represents the functional meaning of a word better than the window based methods (Levy and Goldberg, 2014a).

These count-based vector spaces are extremely large, and thus computationally cumbersome to work with, so one enhancement to them is to run singular value decomposition (SVD) on the vector space, to reduce the dimensionality of each word vector to a reasonable number, usually less than 1000 (Turney and Pantel, 2010).

After learning the vector representation, several tasks are used to evaluate the intrinsic quality of the representation. One example is answering questions found on standardized tests like the TOEFL or SAT. In these tests, question are often of the format "Choose the word that is most synonymous with x" where x is any word in the vocabulary. By determining the distance between the vector for x and all other words in the vector space (or the options given if the question is multiple choice), we

can return the closest one as the synonym (Turney and Pantel, 2010).

Besides querying the vector space, the distributional representation of meaning lends itself very well to machine learning techniques. Clustering can be used to automatically build thesauruses, while classification can be used to predict sentiment. For an extensive list of applications, see Turney and Pantel (2010)

Recently neural network inspired methods have become the preferred method to learn vector representations, one of the most famous being word2vec (Mikolov et al., 2013). Word2vec reformulates the representation learning problem as one of prediction, learning a vector in a low dimensional space that can best predict the context words surrounding the target word. Initially there was much excitement about the improved performance of these methods on standard tasks, as well as the analogy task which was part of the original presentation of word2vec.

As these methods were further inspected, it was shown that word2vec is roughly equivalent to a matrix factorization method, like SVD, and that the values of the hyperparameters are more important to performance on a given task than the actual method used to construct the vector (Levy and Goldberg, 2014b; Levy et al., 2015). None the less, these new approaches have a distinct advantage in that they tend to be very fast to train, and do not ever need to hold a $|V|$ by $|V|$ sized matrix in memory.

Building off vector space semantics, Baroni and Zamparelli propose that adjectives are in fact matrices, if nouns are represented in a standard vector form (Baroni and Zamparelli, 2010). The adjective matrices are multiplied with a noun vector to get the corresponding vector representing the adjective noun phrase. This

representation comes from the view that adjectives are functions between nouns, as proposed by Kennedy (Kennedy, 1997).

2.5.2 Knowledge Based Semantics

The representation of words as vectors is far from the only type of representation found in the literature. Raskin and Nirenburg take an ontological approach in which they give adjectives and their associated properties in the ontology the same standings as nouns and the concepts they are associated with (Raskin and Nirenburg, 1995). For example, an adjective like *big* will modify the property SIZE of a given concept by assigning a high numerical value. An adjective like *red* gives a literal value of red for the COLOR property, raising questions about the symbol grounding problem.

Similar to the approach of Raskin and Nirenburg, the FrameNet project represents lexical semantics as frames, based on the linguistic tradition of frame semantics (Baker et al., 1998). In FrameNet the word *big* invokes the SIZE frame, which has as the value of its ENTITY slot the noun being described. FrameNet does not represent the value of SIZE directly, but does provide definitions for each word that could invoke that frame, such as "large in size" for big.

Another knowledge based representation is found in WordNet. The meaning of an individual word in WordNet is represented by the synonyms of that word, a group called a synset, and the relationships between that synset and other synsets in the lexical network.

For the representation of adjectives, the dumbbell structure described in (Gross

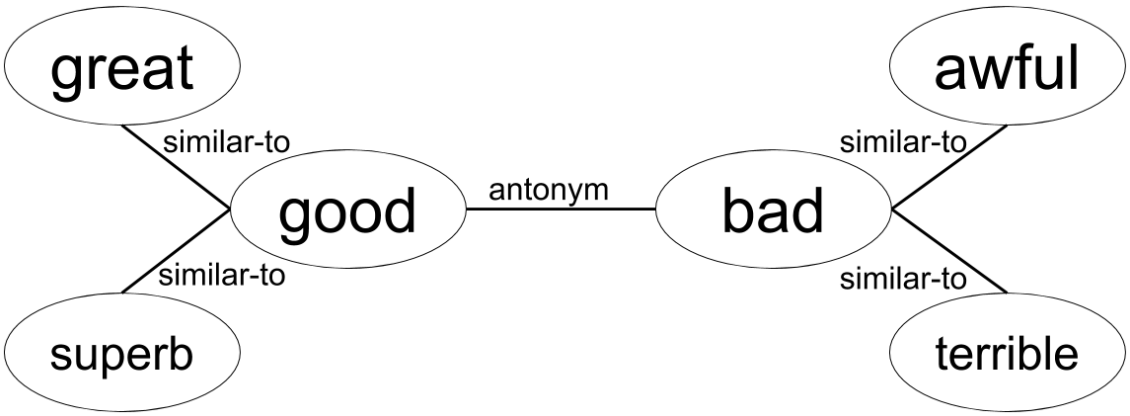

Figure 2.1: Wordnet structure for adjectives of QUALITY

et al., 1989) is used. The central synsets are linked by the antonymy relation, for example *hot* and *cold*. From there, the similar-to relation connects the central synsets to the less central ones, e.g., *warm* is similar-to *hot*. This structure can be seen in figure 2.1. Gross, et al. hypothesize that there are most likely other relations between adjectives, but do not elaborate.

For organization within the lexicon, antonymous adjective synsets are linked with a head noun that names the property it modifies. The authors specifically claim only 2% of adjectives could be represented using a scalar scheme, so their dumbbell representation is more universal. This approach has its faults as well however. For example, among the adjectives noted as similar to *large* are *ample, deep, double, extensive* and *puffy*, yet these should hardly be considered synonyms among themselves (Gross and Miller, 1990).

Becoming even further unstructured, the approach of Abdullah and Frost maintains that adjectives should be represented as sets of the nouns they modify. The examples given show that while this works for many logic based tasks, there is

no room in this representation for relationships between adjectives (Abdullah and Frost, 2005).

2.6 Previous Approaches to Constructing Scales

The creation of adjective scales can be broken down into two components: determining which words make up the scale, and ordering the words on the scale. Prior work usually focuses on one subtask or the other, with a few exceptions.

2.6.1 Scale Membership

The earliest work in NLP concerning adjective scales was done by Hatzivassiloglou and McKeown and concerned determining scale membership. They approach this problem by using two distance measures between pairs of words which are then fed to a clustering algorithm (Hatzivassiloglou and McKeown, 1993) Their first measure quantifies the similarity of the distributions of nouns that the adjectives modify. This is calculated as Kendall's τ between two adjectives' vectors whose features are noun co-occurences. While intuitively this makes sense, it has been shown in (Muehleisen, 1997) that two related adjectives may have vastly different distributional patterns in regards to nouns. Their second measure is that any two adjectives that appear adjacent to each other in a given corpus cannot be in the same group. Under their assumption that the language being investigated is formal written English as found in newswire, this is valid, but as NLP has expanded beyond newswire copora, this assumption is no longer valid. One example of a violation of

ADJ1 if not ADJ2	ADJ1 and perhaps ADJ2
ADJ1 but not ADJ2	between ADJ1 and ADJ2
from ADJ1 to ADJ2	ADJ1 or at least ADJ2

Table 2.1: Patterns used in (van Miltenburg, 2015)

this assumption is the phrase *big huge*. This system had a maximum F-measure of 48.00% when clustering 21 adjectives into 9 clusters.

Taking inspiration from Hatzivassiloglou and McKeown, Shivade, et al. also use clustering to determine scale membership. Rather than calculating distances using linguistic information, they use word2vec vectors for their adjectives, and k-means clustering (Shivade et al., 2015). Based on a Mechanical Turk evaluation in which informants were asked which words do not belong, they achieved an accuracy of 74.36%. In this work, all words input to the clustering algorithm are assumed to lie on a scale. In contrast to these distributional techniques, van Miltenburg uses the six patterns shown in table 2.1 to find pairs of words on the same scale, which he terms scalemates. The pairs of words extracted are checked for semantic similarity using a number of different methods, but the most effective one was the requirement that both words belong to the same entry in the Moby theasaurus (van Miltenburg, 2015). Van Miltenburg was interested in the ordering of these words and no evaluation of the identification phase was carried out.

In all three previous approaches, the scales used were what me might term half-scales. For example, {*big*,*large*,*huge*} would be considered a separate scale from {*small*,*little*,*tiny*}.

Besides work on explicitly learning scale membership, work which links an

adjective to the property it modifies is also relevant. One approach to what is sometimes termed the attribute selection problem is found in (Hartung and Frank, 2010). Vectors are built for both nouns and adjectives using lexico-syntatic patterns. Using the pattern "ATTR of DT? NN is—was JJ", if the pair of *color* for ATTR and *blue* for JJ was found, the entry in the vector for *blue* at the feature *color* would be incremented by 1. An example pattern for noun-attrbute pair extraction is "DT NNs RB? JJ? ATTR", which matches a sentence like "The dog's large size".

This can be looked at as splitting a standard (object, property, value) triple into binary relations. Using ideas from compositional semantics such as vector addition and multiplication, they combine the two vectors to determine the most likely dimension referred to. In their implementation the attribute dimensions are predefined as a handcrafted list of 10 attributes. This work was extended to use topic model based vectors rather that count vectors in (Hartung and Frank, 2011).

Kanzaki, et al. use lexico-syntactic patterns to extract concepts that are associated with the property an adjective modifies. They manually inspect the output of these patterns before using the extracted concepts and a self organizing map to construct a hierarchy of the properties, which can be viewed as a hierarchy of adjectives as well (Kanzaki et al., 2006).

2.6.2 Scale Ordering

We now turn our attention to ordering adjectives on the scale. Many of these techniques assume that the knowledge contained in resources like WordNet

Intense Patterns	Mild Patterns
x even y	y very x
x if not y	not y but x enough
x almost y	y unbelievably x
x no y	y not even x
x perhaps y	y but still very x
extremely x y	
is x but not y	
are x but not y	
are very x y	
is very x y	
x sometimes y	

Table 2.2: Patterns used in (Sheinman and Tokunaga, 2009)

Weak-Strong Patterns	Strong-Weak Patterns
x(,) but not y	not x(,) just y
x(,) if not y	not x(,) but just y
x(,) although not y	not x(,) still y
x(,) though not y	not x(,) bus still y
x(,) (and/or) even y	not x(,) although still y
x(,) (and/or) almost y	not x(,) thought still y
not only x but y	x(,) or very y
not just x but y	

Table 2.3: Patterns used in (de Melo and Bansal, 2013)

or FrameNet is valid for populating scales. While FrameNet has less words per frame and thus is more likely to lead to appropriate scale members, we agree with the arguments of van Miltenburg that WordNet is not suitable for determining membership.

The methods to order adjectives fall into two general styles: pattern based and distributional. The most important work on adjective ordering using patterns is AdjScales. Using WordNet as the source of adjectives, they apply lexico-syntactic patterns like "x if not y", where y is the stronger adjective (Sheinman and Tokunaga,

2009). A full list of patterns used by Sheinman and Tokunaga is found in table 2.2. AdjScales take two or more words and locates them in WordNet, using the similar-to relation to identify what other words should be on the scale. The selected words are then divided into two subscales using the antonym relation in WordNet. The words that participate in the antonymy relation directly are denoted as head words.

For each subscale, the words in that scale are substituted into the patterns along with the head word. For example, if the head word was *big* and another word in the subscale was *huge*, two strings would be generated using the pattern "x if not y", "huge if not big" and "big if not huge". A search engine is then queried with each of these patterns, and the relative intensity of *huge* is incremented by one if "big if not huge" returns 3 times more hits than "huge if not big" and is present in a certain number of documents. After performing this comparison for all words and all patterns, the words are split, binary search style, into words more intense than the head word, and words less intense.

The same procedure is followed in each of the smaller groups of words, except this time in a pairwise fashion among all words. This splitting and ranking of words is continued until the entire subscale is ordered. The final step, which is suggested as optional, is to unify the two subscales. This is done by reversing the order of the subscale with the less frequent head word, and appending it to the left side of the other subscale.

The scales were evaluated by presenting annotators with the subscale groups, and asking them to classify each word as more or less intense than the given headword. The average agreement with the annotators was 86.11% for words predicted to be

weaker than the head word and 70.20% for words predicted to be more intense than the head word. It is important to point out that the evaluation only covers the relationship to the headword, that is the first pass through the data, and does not comment on performance between the other words.

It is our experience, and has been reported by Ruppenhofer, et al. (2014), that the AdjScales method has very low recall in terms of the number of adjective pair-pattern combinations that successfully return hits in a search engine. One method suggested to overcome this is to use pairwise data between all adjectives in a subscale, rather than in a binary search like paradigm.

De Melo and Bansal use Mixed Integer Linear Programming to accomplish this, using patterns similar to but not identical to those of Sheinman and Tokunaga (De Melo and Bansal, 2013). The patterns used by DeMelo and Bansal are shown in table 2.3, and rather than being divided into strong and mild patterns, they are divided into strong-weak patterns, which show the first word is greater than the second, and weak-strong patterns, which are evidence that the first word is less intense than the second word. These patterns are used to search the Google Ngrams corpus for counts.

From these counts, a score is calculated between each pair of words in the subscale, scaled by the frequency of the individual words. The formula is shown in equations 2.1-2.7, where a positive score indicates word a_1 is to the left of word a_2 and a negative score indicates the reverse. These scores are used as input to a MILP, whose objective is to maximize the score of all pairs of words, multiplied by the distance between those words on the scale. The complete definition of the mixed

integer linear program is shown in equation 2.8

$$P_1 = \sum_{p_1 \in P_{\text{weak-strong}}} \text{count}(p_1) \tag{2.1}$$

$$P_2 = \sum_{p_2 \in P_{\text{strong-weak}}} \text{count}(p_2) \tag{2.2}$$

$$W_1 = \frac{1}{P_1} \sum_{p_1 \in P_{\text{weak-strong}}} \text{count}(p_1(a_1, a_2)) \tag{2.3}$$

$$W_2 = \frac{1}{P_1} \sum_{p_1 \in P_{\text{weak-strong}}} \text{count}(p_1(a_2, a_1)) \tag{2.4}$$

$$S_1 = \frac{1}{P_2} \sum_{p_2 \in P_{\text{strong-weak}}} \text{count}(p_2(a_1, a_2)) \tag{2.5}$$

$$S_2 = \frac{1}{P_2} \sum_{p_2 \in P_{\text{strong-weak}}} \text{count}(p_2(a_2, a_1)) \tag{2.6}$$

$$score(a_1, a_2) = \frac{(W_1 - S_1) - (W_2 - S_2)}{\text{count}(a_1) \cdot \text{count}(a_2)} \tag{2.7}$$

$$
\begin{aligned}
&\text{maximize} \quad \textstyle\sum_{(i,j)\notin E}(w_{ij} - s_{ij}) \cdot score(a_i, a_j) - \sum_{(i,j)\in E}(w_{ij} + s_{ij})C \\
&\text{subject to} \\
&\qquad d_{ij} = x_j - x_i && \forall i, j \in \{1, \ldots, N\} \\
&\qquad d_{ij} - w_{ij}C \leq 0 && \forall i, j \in \{1, \ldots, N\} \\
&\qquad d_{ij} + (1 - w_{ij})C \geq 0 && \forall i, j \in \{1, \ldots, N\} \\
&\qquad d_{ij} + s_{ij}C \geq 0 && \forall i, j \in \{1, \ldots, N\} \\
&\qquad d_{ij} - (1 - s_{ij})C \leq 0 && \forall i, j \in \{1, \ldots, N\} \\
&\qquad x_i \in [0, 1] && \forall i \in \{1, \ldots, N\} \\
&\qquad w_{ij} \in \{0, 1\} && \forall i, j \in \{1, \ldots, N\} \\
&\qquad s_{ij} \in \{0, 1\} && \forall i, j \in \{1, \ldots, N\}
\end{aligned}
\tag{2.8}
$$

They evaluate their method against 88 subscales, which were all derived from WordNet. They asked annotators to order the words in each subscale, and compared this to the output of the MILP, using both pairwise accuracy and Kendall's τ. They achieved a pairwise accuracy of 69.6% and a τ of 0.57.

Van Miltenburg use the patterns shown in table 2.1 to order adjective pairs, keeping the most frequent instantiation of the pattern as evidence. These produce up to 32,470 pairs of adjectives, although only 2,611 of those pairs exist in the evaluation dataset. The highest performing similarity filter returns 2,230 pairs, of which 287 are in the evaluation dataset. For evaluation, van Miltenburg compares the ordering of a pair to the subjectivity lexicon released as part of the Multi-Perspcetive Question Answering Project (MPQA) (Wilson et al., 2005). Words in the MPQA lexicon are assigned to either the "weakly subjective" or "strongly subjective" classes, and if

the ordering predicted by the pattern produces a weakly suggestive word followed by a strongly subjective word, it is counted as correct. Using this methodology, the top performing parameter setup produces a score 63.86%.

Using a system similar to the one described by Nirenburg and Raskin (1995), Hickman, et al. (2015) are interested in learning the ontological representation of adjectives. The most relevant piece of this system to our work is what they term the "Compare Objects" step, in which they determine the value of an unknown adjective by finding its use in a corpus along with a numerical value. The example given is that a "tiny tract of land" was earlier referred to in an article as being 31 acres. Later in that same article, it is established that the same referent is located on a 30,000 acre land mass. From this they determine that *tiny* represents the proportion of 31:30,000, at least in one example.

It is not clear if this work is automated, or a proof of concept that was done manually. Only adjectives of size were investigated and no evaluation was conducted beyond commenting on the fact that *tiny* is in fact on the low end of the size scale, and thus a ratio of 31:30,000 is logical.

While pattern based methods have been the norm, recently work using distributional semantics has been done. Expanding on the vector space algebra introduced by Mikilov (2013) to solve the analogy problem, Kim & de Marneffe present equation 2.9 as a way to determine which word is the midpoint between two given words a and b. They extend this to calculate the points at each quartile along the line between two words in vector space to produce scales, although neither these scales nor the midpoints are ever directly evaluated. Instead, the effectiveness of this formulation

is evaluated on the performance on the Indirect Question Answer Pair (IQAP) task.

$$w_b + \frac{w_a w_b}{2} \tag{2.9}$$

In this task, the system is presented with a question answer pair as shown in example 2.1 below and must infer if the respondent should be interpreted as having said "yes" or "no". To use their equation to answer this question, they find the antonym of the adjective in the question, and then evaluate if the adjective in the answer is on the side of the halfway point between the adjective in the question and its antonym or not. They achieve an F1 score of 70.58 using this solution.

(2.1) Question: Do you think that's a good idea?

Answer: It's a terrible idea.

Using the conclusions of Kennedy (2005) and Paradis (2001) that adjectives more towards the end of a scale tend to occur with a different class of adverbs than those more central, Ruppenhofer, et al. (2014) use an association score between the adjectives in question and two sets of adverbs to determine which adjective is more extreme. They evaluate their scales against gold standard scales constructed by annotators by using Spearman's ρ. These scales are not the type we have been discussing, but can be better thought of as buckets, such as a bucket for "very high positive","high positive", etc. According to this metric, their collocation based approach performs well across four scales. They compare their system to several other existing methods, including those based on sentiment lexicons, and critically, demonstrate that while sentiment is equivalent to scale position for some scales, it

has very little correlation to other scales, such as the scale for size.

One sentiment based method is that of de Marneffe, et al (2010). They use IMDB review data to calculate the probability of a rating given a word. From this they estimate the scale position of each word. This paper introduces the IQAP task previously discussed.

2.7 Lexical Substitution

Unrelated to the problem of learning scales prior to this dissertation, we will now review the lexical substitution task and the relevant approaches to solving it.

The lexical substitution task was proposed as a variation of the word sense disambiguation task in a more realistic environment (McCarthy and Navigli, 2007). Rather than trying to select the proper sense of a word directly, which would require a sense inventory, the goal of lexical substitution is to select the best possible substitute for a word, given its context. An example from the Senseval 2007 test data set is "The chain will only be as **strong** as its weakest link." Some of possible substitutes for *strong* are *heavy*, *capable*, *tough*, *secure*, *resolute*, and *powerful*

We will refer to the word being substituted, in this example, *strong*, as the target word. We will refer to the possible substitution words as candidate words, and the words surrounding the target word as context words.

While Lexical Substitution predates the recent resurgence of vector based semantics, it has been used to investigate the modeling of context and word similarity. Melamud first solved this task using a vector space model by saving the context

vectors produced in word2vec and defined a series of substitutability metrics based on the cosine distance between the target word and the candidate as well as the candidate and the context (Melamud et al., 2015). The proposed metrics are given in equations 2.10 - 2.13.

$$\text{Add} = \frac{\cos(s,t) + \sum_{c \in C} \cos(s,c)}{|C| + 1} \tag{2.10}$$

$$\text{BalAdd} = \frac{|C| \cos(s,t) + \sum_{c \in C} \cos(s,c)}{2|C|} \tag{2.11}$$

$$\text{Mult} = \sqrt[|C|+1]{\frac{cos(s,t)+1}{2} \prod_{c \in C} \frac{cos(s,c)+1}{2}} \tag{2.12}$$

$$\text{BalMult} = \sqrt[2|C|]{\left(\frac{cos(s,t)+1}{2}\right)^{|C|} \prod_{c \in C} \frac{cos(s,c)+1}{2}} \tag{2.13}$$

Roller et al. proposed a new formula using the same vectors, which estimates the conditional probability of a substitute given its context, $P(s|C)$, and the conditional probability of a substitute given the target word, $P(s|t)$ (Roller and Erk, 2016). These conditional probabilities are estimated using a softmax function over the dot products of all substitute word vectors with the target vector and context vectors respectively. Roller et al. present two models, PIC and nPIC. PIC learns a weight matrix to weight the context vector, while nPIC has no parameters. PIC outperforms nPIC slightly, but both outperform the Add methods introduced by Melamud. The equations for PIC and nPIC are shown in equations 2.14 and 2.15, where Z_t, Z_C, and Z_n are normalizing values to ensure the value is between 0 and 1.

$$\text{PIC}(s|t,c) = \frac{1}{Z_t}\exp\left\{s^{\mathrm{T}}t\right\} \times \frac{1}{Z_C}\exp\left\{\sum_{c\in C} s^{\mathrm{T}}[Wc+b]\right\} \tag{2.14}$$

$$\text{nPIC}(s|t,c) = \frac{1}{Z_t}\exp\left\{s^{\mathrm{T}}t\right\} \times \frac{1}{Z_n}\exp\left\{\sum_{c\in C} s^{\mathrm{T}}c\right\} \tag{2.15}$$

Melamud introduced context2vec a LSTM based method that is trained specifically to perform the lexical substitution task (Melamud et al., 2016). The objective function is the same as in word2vec, but rather than the context being a single word, the context is the representation calculated by the LSTM RNN. By using LSTM's, context2vec is able to consider the whole sentence as the context rather than just a window or dependencies. Context2vec achieves state-of-the-art results or near state-of-the-art results on standard lexical substitution tasks.

2.8 Summary

In this chapter we have reviewed the linguistic literature about adjectives and scales, highlighting the particular interpretations of scales that we will use through out this dissertation. We next reviewed both the distributional and knowledge based methods for representation of lexical semantics computationally. Following this we presented the previous approaches to solving both the scale membership and scale ordering portions of our problem. Finally we introduce the lexical substitution task, and three approaches to solving it that will be used in our solution to the problem of learning scales.

Chapter 3: Data

3.1 Introduction

Before turning our attention to solutions for identifying and ordering adjective scales, we must address a significant obstacle. There is very little data available concerning adjective scales, especially their membership. In this chapter we discuss two datasets that we built, a gold standard of scales that was originally presented at LREC (Wilkinson and Oates, 2016), and an expanded dataset for the indirect question answer task.

3.2 Gold Standard

In this section we present a gold standard of 12 adjective scales for use in evaluation of methods to identify and order adjective scales as well as for use in investigating scalar implicature, a need highlighted by Van Tiel et al (2016) [1]. We use cultural consensus theory (CCT) to both produce the gold standard as well as to gain insight on the level of consensus among the informants (Romney et al., 1986).

CCT was developed to aggregate the shared knowledge of a domain by a culture

[1] Available at https://github.com/Coral-Lab/scales

(Weller, 2007). It has roots in test theory and was developed as an analysis of latent variables of participants that can be done when the true answers are unknown as opposed to other methods such as Classical Test Theory or Item Response Theory (Batchelder and Romney, 1988). This provides a useful framework to judge an informant's understanding of the task without predetermining what words should be on the scale or not, and to use the informant's competency when constructing the standard.

The members of a scale are collected through free-listing, an elicitation method in which informants are asked to list as many words, phrases, or ideas they can think of in response to a prompt (Weller and Romney, 1988). While CCT has been applied to data gathered through free-listing in the past, it was primarily used to determine if consensus existed in a culture about a topic. We believe we are the first to determine the culturally salient answers through CCT with data elicited using free-listing. We do this through the use of the bias variable available in CCT. After determining the salient members of each scale, we ask informants to order the sets of words as best they can. In the second task we again use CCT to produce the ordering. The high-level relation between the two phases is shown in figure 3.1. The details of each phase, shown inside the dashed lines, will be discussed in sections 3.2.2 and 3.2.3.

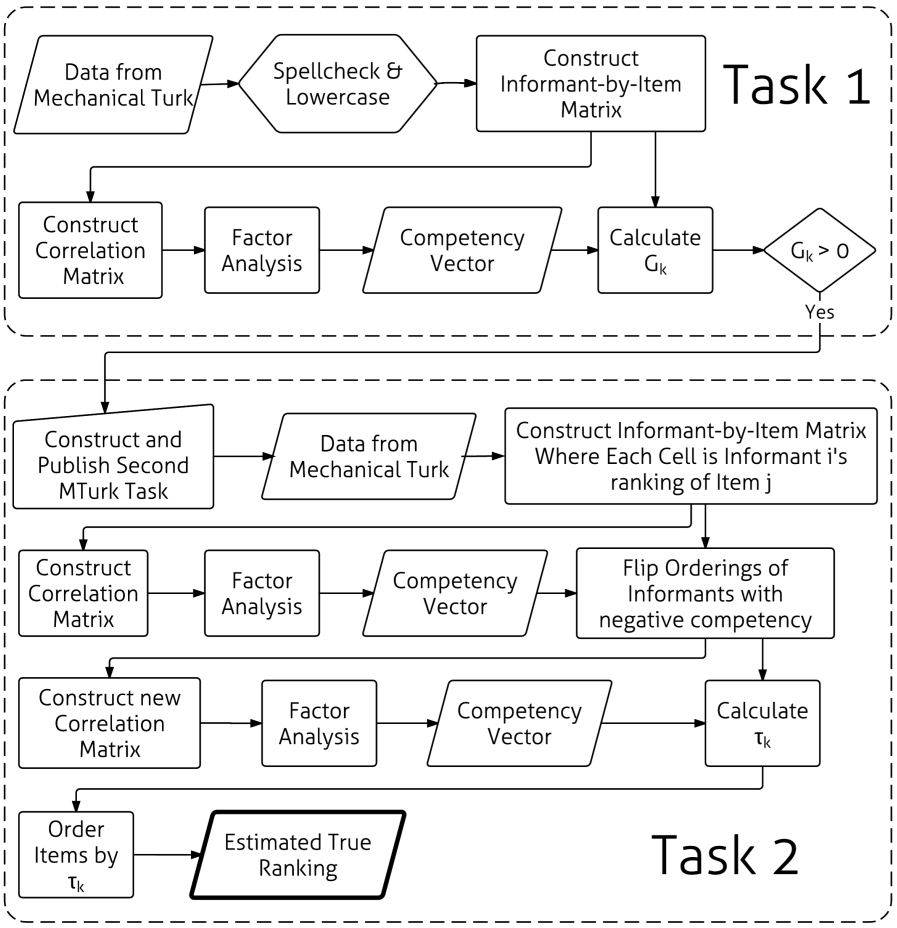

Figure 3.1: Overview of methodology.

3.2.1 Previous Datasets of Scalar Adjectives

As part of their work creating adjective scales, discussed in section 2.6.2, Ruppenhofer et al. proposed a gold standard of adjective orderings derived from pairwise comparisons of words on a scale. (2014). The words in this study all belong to the same frame in FrameNet. The words were then grouped into buckets, based on whether the majority of informants rated word_1 higher than word_2, word_2 higher than word_1, or word_1 "as intense as" word_2 (Ruppenhofer et al., 2014). A gold

standard for 4 scales was produced this way. This paper differs from our work in two ways: we collect the sets of words to be ordered empirically, and we produce a total ordering of words.

Our work is inspired by Sutrop's (1998) study of Estonian temperature words, which consists of first determining the words for temperature and then ordering them. The methodology was slightly different than ours as each informant ordered all of the words they themselves provided for temperature. In contrast, we collect a list of words as one task, and then after performing aggregation, present informants with the same words to order as a separate task.

3.2.2 A Gold Standard for Scale Membership

The first standard we produce is an empirically determined group of adjectives that are members of the same scale. This is important because, as De Melo (2013) notes, existing resources are often more broad in their groupings than what is acceptable for a single scale. While work on scale membership has been limited, several taxonomies of adjective groupings have been proposed. To cover a variety of adjective types, we use Dixon's typology (1977) as a guide in choosing which scales to include in the dataset. Dixon proposes 7 semantic classes of adjectives: DIMENSION, PHYSICAL PROPERTY, COLOR, HUMAN PROPENSITY, AGE, VALUE, and SPEED. The groups not only have a common semantics and semantic opposition behavior, but also similar morpho-syntactic behaviors (see Table 1 of (Dixon, 1977)).

The majority of adjectives in English belong to either the PHYSICAL PROPERTY

or HUMAN PROPENSITY groups according to Dixon. Since the scalarity of color words is unclear (Bolinger, 1977), we exclude scales of that type and chose one scale from each of the remaining types to investigate, adding additional scales for the two groupings Dixon lists as more common. Dixon also notes that the words *easy* and *difficult* do not fit neatly into this system, yet to our intuition are scalar, and thus were included in this study as well. Finally, although not typically viewed as adjectives in current linguistic thought, quantifiers such as *few*, *some*, and *many* are among the most commonly studied scalar items and were included as well. All together this gives us 12 scales to investigate.

3.2.2.1 Methodology

Ideally, an informant would be asked which words belong on a scale directly, using a prompt such as "List all adjectives that describe an object's temperature", as was done by by Sutrop (1998). While words like *temperature* or *intelligence* succinctly describe a scale, many scales exist that do not have this luxury. For example, it is difficult to think of a single word that describes a scale containing *big* and *small* but not *tall* or *wide*. Rather than attempt to identify each scale with a noun, we instead use prompt words we believe could be on the scale. These words were chosen from lists of synonyms and antonyms as provided by dictionaries and thesauruses as a proxy to naming the scale.

Given a set of prompt words that are hypothesized to be members of a scale, we present the informant with three of the words, randomly chosen. To ensure that

the prompt was representative of the entire scale, all three words were not permitted to be from the same side of the scale. For example, if the set of prompt words was {*large*, *huge*, *colossal*, *small*, *tiny*, *microscopic*}, we would not want the informant presented with the first three. This variation was ensured by splitting the set of possible prompt words into two groups of synonyms or near synonyms based on existing resources. The prompt was constructed by randomly picking two words, one word from each group and then randomly picking the third word from the remaining words in both groups. The three words were then shuffled.

It is important to note that this task is solely focused on eliciting scale membership. The existing resources were used only to construct prompts and are not taken as truth. CCT determines an informant's competence without regard to a prior established truth. A method that avoids this intervention by the researcher would unquestionably be superior however, and further research is needed on this.

Once the prompts are selected, the informant was then asked to list all the other adjectives they felt were similar to the three listed adjectives. This question was repeated for all 12 postulated scales. In addition the informants were presented with the same question with 4 groups of adjectives that were not believed to form a scale. Some of these control groupings contained related words, such as adjectives describing material, while others were unrelated.

All questions were presented on a single page, with each informant seeing the questions in a random order. This task was given to 500 informants on Amazon Mechanical Turk (AMT) who were paid 50 cents for their participation. This task was available to all members of AMT with no requirements. An example presentation

of this task is shown in figure 3.2.

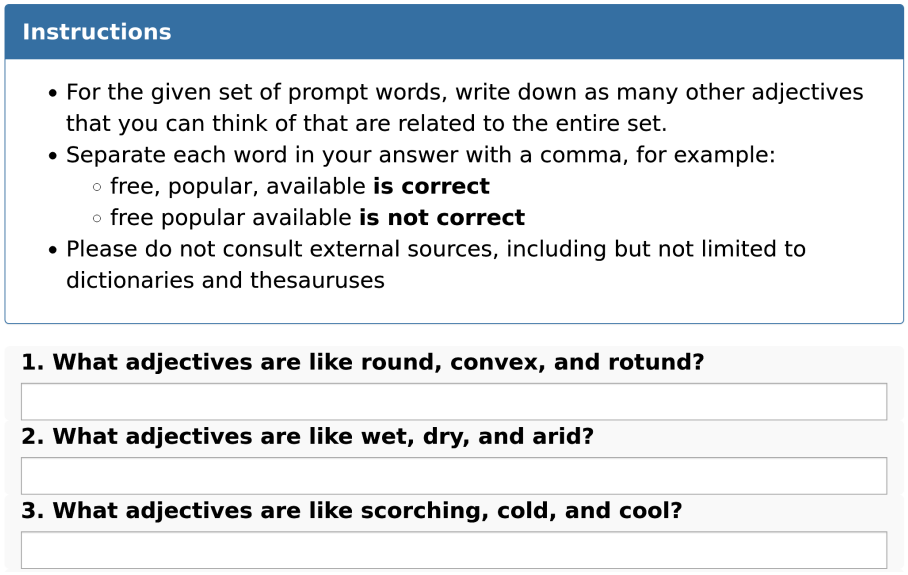

Figure 3.2: Lexical elicitation interface.

3.2.2.2 Results

The study was completed in 97 hours and 35 minutes and the average response time was 9 minutes 17 seconds. The average response length was 3.098 words with a standard deviation of 0.354 words over the 16 sets of words. The average number of words in an answer given a scale ranged from 2.7 for a random selection of adjective prompt words {*acidic*, *magnetic*, *supersonic*, *savory*, *overpriced*, *classified*} to 3.888 for words about SIZE.

We used Cultural Consensus Theory (CCT), a framework pioneered by Romney, Weller, and Batchelder (1986), to determine the shared belief of scale membership. Given that the data was open ended we used the informal variant.

In this variant, each informant's response is transformed into a vector over all the responses for a set of prompt words, placing a one in the column if they mentioned the word, and a zero otherwise. To standardize the data we ran spelling correction from hunspell[2] on each word and accepted the first alternative spelling in all cases where hunspell indicated a misspelled word. CCT can be broken into two steps, calculating the competencies of informants and determining if a consensus exists, and using the competencies and responses to produce the correct answers.

In traditional CCT an informant-by-informant correlation matrix is created and then factor analysis is run on the matrix. Due to variation in prompt words, we made the following change. When comparing two informants, if one informant listed a word and the other informant was given that word as a prompt word, the second informant was assumed to have included it. If both are given a prompt word, neither are assumed to have included it. This ensures that informants were not penalized for not listing their prompt words, but at the same time are not rewarded for having the same prompt word as another informant. See figure 3.3 for a visual explanation of this, where a one in a vector indicates an informant responded with that word and a zero indicates they did not.

After running factor analysis on the matrix, the first factor gives the competencies of the informants and the ratio between the first and second eigenvalues provides insight into the amount of consensus. The generally accepted ratio that indicates consensus is 3:1 (Weller, 2007). The eigenvalue ratios for the 16 groups of words are presented in table 3.1. Given the competencies, the estimated true answers can be

[2]http://hunspell.sourceforge.net/

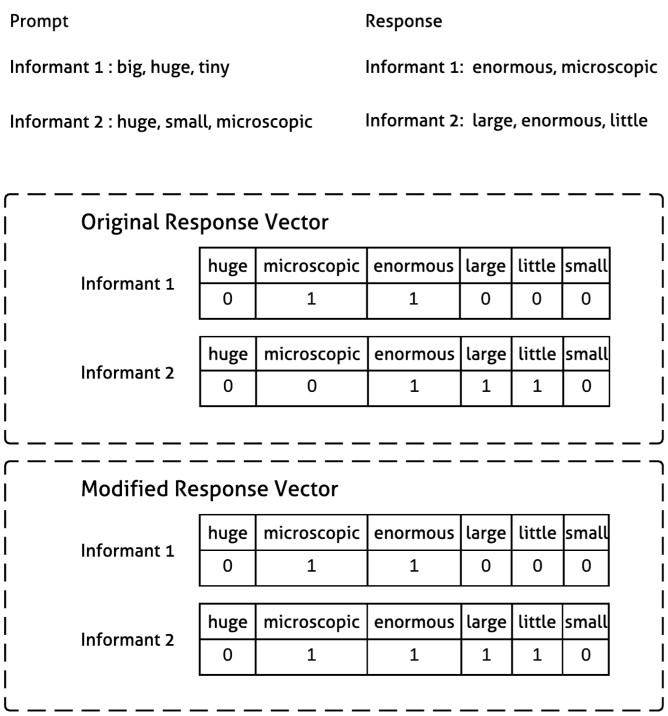

Figure 3.3: Comparing informants with different seed words

When informants 1 and 2 are compared, informant 2 is assumed to have included *microscopic* for this comparison only as informant 2 did not have the opportunity to list it. Note neither informant is assumed to have included *huge*, although both had it as a prompt.

calculated using equation 3.1. This variation of the calculation was first introduced by Batchelder and Romney (1988). A positive value for G_k represents a shared belief that word k is part of the scale. Here we are evaluating a single potential word, indexed with k. X_{ik} is the ith informant's response, D_i is their competency, and g is the bias. The bias was originally intended to model each informant's bias in response to the question when guessing. We set the bias to be the average response length for a question divided by the number of words given in responses (the length

Sample Words	Eigenvalue Ratio
smart, dumb, stupid	9.26
ugly, beautiful, gorgeous	8.45
hot, cold, freezing	7.67
old, new, ancient	7.46
fast, quick, slow	6.71
same, different, similar	6.68
many, few, some	6.63
tiny, big, huge	6.49
easy, hard, simple	6.15
wet, dry, damp	5.87
terrible, great ,bad	5.14
bright, dark, light	4.05
round, circular, concave	4.91
skinny, fat, hairless	2.23
plastic, wooden, metal	2.56
expensive, secret, attractive	1.64

Table 3.1: Eigenvalue ratios for 16 sets of words, proposed scales above the line and sets of adjectives that do not make up scales and were used as controls below the line.

of the response vector). This can be viewed as a heuristic of the informant deciding when to stop listing items.

$$G_k = \sum_{i=1}^{N} X_{ik} \ln \frac{(D_i(1-D_i)g)(1-(1-D_i)g)}{(1-D_i)^2 g(1-g)} - \ln \frac{1-(1-D_i)g}{(1-D_i)(1-g)} \tag{3.1}$$

The inspiration for the use of the bias variable was due to an observation that out of more than 100 unique words elicited for each scale, most informants list only three or four of them. The lack of a mention cannot be taken solely as evidence that the informant believes that word is not in the set. Mechanical Turk informants are trying to make money, and may spend less time on a task, so there is ambiguity in

whether a zero in the response vector indicates a given word doesn't belong, or that the informant simply didn't think of it while rushing through.

Because we are using informal CCT, and not asking the actual question of whether a word belongs in a set, some competencies were slightly over 1.0. These were set to .999. The results for three scales are visible in table 3.2. A positive G_k indicates that there is cultural consensus about the word, while a negative G_k indicates that the word is not a consensus member of the scale. The words in bold are the gold standard members of their respective scales.

Word	G_k
∗**tiny**	**856.82**
big	**601.23**
∗**huge**	**561.04**
∗**small**	**527.43**
gigantic	**421.08**
∗**large**	**164.20**
minuscule	**116.97**
enormous	**34.70**
∗microscopic	-85.32
little	-87.232
giant	-200.20
∗colossal	-226.70
micro	-242.83
gargantuan	-268.18
massive	-281.17

(a)

Word	G_k
∗**easy**	**929.44**
∗**hard**	**771.91**
simple	**718.75**
∗**difficult**	**684.37**
∗effortless	-126.80
challenging	-184.52
effort	-274.10
tough	-276.670
∗painless	-303.56
∗herculean	-344.53
strong	-397.42
impossible	-444.99
painful	-465.17
complex	-560.39
arduous	-562.64

(b)

Word	G_k
∗**plastic**	**167.39**
∗**wooden**	**152.56**
metal	**130.58**
hard	**100.05**
∗**glass**	**9.90**
∗**stone**	**3.71**
∗metallic	-5.79
wood	-22.91
solid	-55.80
∗concrete	-80.11
brick	-119.13
rock	-133.08
ceramic	-148.91
cement	-152.61
shiny	-167.99

(c)

Table 3.2: G_k for words along two postulated scales (a, b) and one set of adjectives that describe material (c). Words marked with ∗ were prompt words.

3.2.2.3 Toy Example

To assist in understanding this process we will walk through a toy example. Suppose 4 informants are asked what adjectives they feel go with various prompt

words for size. We may get responses such as in table 3.3.

Informants	Results
I_1	small, minuscule, tiny, big, huge
I_2	big, large, miniscule
I_3	TINY, LARGE, HUGE
I_4	wrong, bad, other

Table 3.3: Example responses for toy example.

After standardizing the responses by using spell check and converting all words to lowercase, we build an item-by-informant matrix as shown in figure 3.4 and calculate the informant-by-informant correlation matrix as shown in figure 3.5.

	bad	*big*	*huge*	*large*	*minuscule*	*other*	*small*	*tiny*	*wrong*
I_1	0	1	1	0	1	0	1	1	0
I_2	0	1	0	1	1	0	0	0	0
I_3	0	0	1	1	0	0	0	1	0
I_4	1	0	0	0	0	1	0	0	1

Figure 3.4: Item-by-informant matrix.

	I_1	I_2	I_3	I_4
I_1	1	0.16	0.16	−0.79
I_2	0.16	1	0	−0.50
I_3	0.16	0	1	−0.50
I_4	−0.79	−0.50	−0.50	1

Figure 3.5: Informant-by-informant correlation matrix.

The first factor produced by factor analysis on the correlation matrix in figure 3.5 represents the informants' competencies and is shown in figure 3.6. This places a numerical value on the intuition that I_1 lists culturally salient words, while I_4 has

either misunderstood the task completely or is responding maliciously. To find G_k for *big* we apply equation 3.1 to competency vector D and the column labeled *big* in figure 3.4. In this example, g is equal to 3.5/9 or 0.388. When this equation is reduced, G_{big} comes out to be 2.50, indicating that *big* is a member of the scale in the toy example. This procedure is repeated for all other columns in the response matrix.

$$\begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \end{array} \begin{array}{c} D \\ \left[\begin{array}{c} 0.79 \\ 0.498 \\ 0.498 \\ -0.998 \end{array}\right] \end{array}$$

Figure 3.6: Competency Vector.

Having determined the culturally shared belief of scale membership we next evaluate the effect of the prompt words on the output. 65% of the prompt words were deemed salient according to the analysis. Running Fisher's exact test on each word grouping, 14 of the 16 groups have a significant relationship between a word being a prompt word and being part of the shared cultural belief. The two exceptions were the group of words representing generic adjectives about appearance and the group of random adjectives. Further analysis is needed to determine if the significance is due to the words being prompt words or the authors themselves being native English speakers and thus possessing some of the shared belief, thereby influencing the choices of prompts.

3.2.3 A Gold Standard for Scale Ordering

In the second task, informants were asked to order the words in each scale. For this phase of the study, only the 12 adjective scales were used. 200 informants from Mechanical Turk participated and were again paid 50 cents each. For each scale, the words with a positive G_k from the analysis of the first phase were placed into a random order. The informant was asked to drag and drop the words into the order they felt was best. The instructions were intentionally left vague as to not presuppose which end of the scale was higher. In addition, each scale was followed by a text box allowing the users to enter any words that they felt did not belong in the group. The 12 scales were randomly shuffled for each informant. This interface can be seen in figure 3.7.

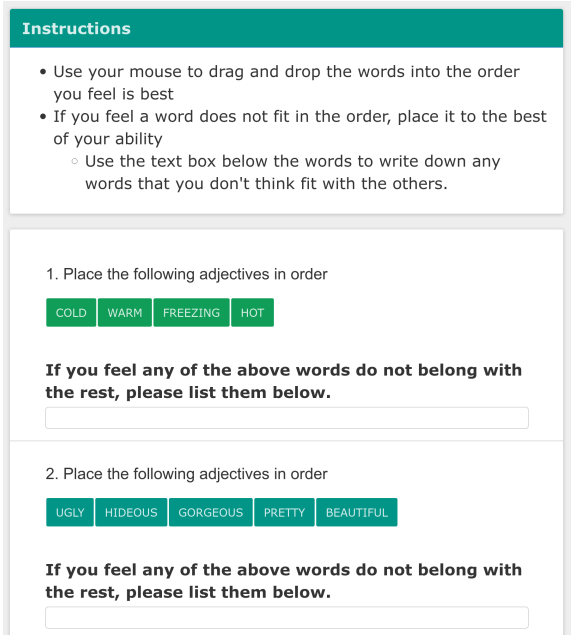

Figure 3.7: Adjective Ordering Interface.

3.2.3.1 Results

This task was completed in 121 minutes with an average of 11 minutes 19 seconds per informant. We used the formulation of CCT designed to process rank order data, as put forth in (Romney et al., 1987) to analyze the data. If an informant did not attempt to order a particular word set, meaning no words were ever moved, that informant's answer was not used when analyzing that word set.

Given that the instructions were vague, it is not surprising that informants produced orders with different orientations. To avoid researcher bias in determining the orientation, we determined the informants' competencies and then for any informant who had a negative competency for a given scale, we reversed their ordering. This allowed us to orient all scales in the same direction without specifying which direction was positive.

Following this we ran the complete CCT pipeline. Romney, et al. give the formula shown in equation 3.2 for finding the true ordering, where z_{ik} is informant i's normalized rank for word k and τ_k is the score for word k. Each informant's ranking is normalized to have a mean of zero and a standard deviation of one.

Words are then ranked according to their τ_k value. The recommended method for finding β is equation 3.3, where R is the informant-by-informant correlation matrix and r_t is the competency vector. Unfortunately our data resulted in a singular matrix for R. As suggested by Romney, et al. we used the competencies directly as an estimate for β.

$$\tau_k = \sum \beta_i z_{ik} \tag{3.2}$$

$$\beta = R^{-1} r_t \tag{3.3}$$

Scale	Eigenvalue Ratio
minuscule, tiny, small, big, large, huge, enormous, gigantic	29.47
horrible, terrible, awful, bad, good, great, wonderful, awesome	18.68
freezing, cold, warm, hot	15.99
hideous, ugly, pretty, beautiful, gorgeous	12.28
parched, arid, dry, damp, moist, wet	11.87
dark, dim, light, bright	10.78
idiotic, stupid, dumb, smart, intelligent	8.99
ancient, old, fresh, new	7.58
simple, easy, hard, difficult	7.20
few, some, several, many	6.75
same, alike, similar, different	6.60
slow, quick, fast, speedy	3.52

Table 3.4: Scale orderings and the corresponding eigenvalue ratios

Table 3.4 gives the gold standard that can be used for evaluation. Each row gives a scale with it's members ordered and, although no information was provided to the informants on the directionality of the scales, they seem to match our intuition. While all orderings qualify as culturally salient according to the eigenvalue ratio, there is a wide range of consensus. The scales that display high consensus values are among some of the most commonly researched in literature. Table 3.5 shows the τ_k values for three scales in our dataset, those for SIZE, TEMPERATURE, and INTELLIGENCE. The τ_k should be interpreted with caution. While it is tempting to say that *big* and *large* are closer on the scale than *minuscule* and *tiny*, this would need to be independent validated. Part of what τ_k captures is certainty in position,

Word	τ_k
minuscule	-264.45
tiny	-202.88
small	-129.71
big	-4.61
large	24.52
huge	123.96
enormous	222.45
gigantic	230.71

(a)

Word	τ_k
freezing	-223.33
cold	-93.95
warm	94.27
hot	223.01

(b)

Word	τ_k
idiotic	-205.30
stupid	-108.31
dumb	-57.01
smart	146.36
intelligent	224.26

(c)

Table 3.5: τ_k for words along the scales for SIZE (a), TEMPERATURE (b), and INTELLIGENCE (c)

which is why we believe the scale for TEMPERATURE appears so symmetrical. We should not conclude from this that *warm* is actually the correct antonym for *cold*, only that in a list of four words, they are commonly the second and third word.

Looking at the responses to which words should be left out, only 3 words were listed by more than 5% of respondents: *fresh*, *difficult*, and *slow*. *Difficult* and *slow* are both members of four word scales where the other words all represented the positive side. *Fresh* has the lowest G_k of it's scale, but no correlation could be found between the number of informants indicating a word did not belong and it's G_k.

3.2.4 Comparison against other hand created data sets

Ruppenhofer, et al. (2014) construct 4 scales, three of which we also investigate: QUALITY, SIZE, and INTELLIGENCE. Although their standard presents a scale divided into buckets of intensities rather than a strict ordering, we still feel a comparison is warranted. For SIZE adjectives, our ordering reflects their order of intensities, with *gigantic* and *enormous* being labeled as high positive intensity, *big*, *large*, and *huge*

being labeled as medium positive intensity, *small* being labeled as low negative and *tiny* being labeled as medium negative. *Minuscule* was not included in their study as it is not in FrameNet.

All adjectives of INTELLIGENCE in our study were present in theirs and are ordered the same when analyzed in the same fashion as the SIZE scale. FrameNet does not include *horrible*, *terrible*, or *awesome* under the frame for QUALITY. The other adjectives for QUALITY are ordered the same in both studies.

Another comparison we can make is against Sutrop's scale of temperature terms in Estonian. While the methodology is different, Sutrop's final scale in the English equivalents to the original Estonian is ⟨*cold,cool,warm,hot*⟩ while the scale produced from the informants' data in this study is ⟨*freezing,cold,warm,hot*⟩.

3.2.5 Discussion

In this study we have presented the use of Mechanical Turk for elicitation of lexical items rather than just labeling. Our results show that this is a viable resource for lexical elicitation.

This gold standard was designed to favor precision over recall. We aimed not to include every word for a scale but to ensure that the words being ordered by informants are all in fact part of that scale. This dataset can be used to test multiple things. While the most obvious is to test automatic ordering methods, the data can also be used as an additional benchmark for semantic relatedness of word representations. If we take analogies to represent relationships, then we can add

analogies such as "large is to enormous as smart is to _____".

Between the two studies, there was an overlap of 6 informants.

3.2.6 Future work

This study provides a gold standard of adjective orderings, but these ordering are often incomplete. Further work needs to be done on adding more relevant words to each scale. Now that a base collection of words exists for each scale, one extension is to run a study similar to the one used for elicitation of scale membership, but present all informants with the entire known scale in random order and ask what other words belong.

Another important contribution that is needed is to determine how the consensus measurements should be interpreted. From the results discussed above, it is clear that some scales have much more consensus than others, both in the words they include and their ordering. It is an open question if this lack of consensus is due to the scale being more difficult in some sense or if it is an indication that the words given do not constitute a single scale.

One improvement in analysis of the elicitation task is to incorporate list position as is done when calculating the salience index (Sutrop, 2001). Salience index was not used in this work because while it produces a very logical ranking of cultural salience, there is no consistent cut off point on which words to include as part of the scale.

This methodology needs to be replicated with more sets of words and in other

languages. Replication will provide insight into which groups of words do constitute scales, and those that do not. From this data we will be able to determine if the eigenvalue ratio has a different threshold for data gathered by free-listing than the 3:1 ratio used in literature. Replication in other languages will also provide an avenue to investigate the relationship between prompt words and responses by removing researcher bias from being a native speaker of the language.

3.2.7 Conclusion

We have shown that bias term from CCT can not only be used to determine if a scale is culturally salient, but what the salient members of that field are. We have also shown that Amazon Mechanical Turk can be used for lexical elicitation. Furthermore, we have developed a freely available resource for use in both evaluation and linguistic inquiry on scalar adjectives and the scales they create.

3.3 More Indirect Question Answer Pairs

The second dataset we develop is for an application of adjective scales, often used as an extrinsic evaluation of their correctness. One of the most common ways to evaluate adjective scales is to use the Indirect Question Answer Pair (IQAP) task, first described by de Marneffe, et al. (2010). The objective in this task is to predict if the answer to a yes/no question should be interpreted as a *yes* or a *no* when the question and the indirect response are provided. An example from the dataset released with de Marneffe, et al.'s paper is seen in example 3.2. This response should

be interpreted as a *no*, because *terrible* is less than *good* on the scale they share.

(3.2) Question: Do you think that's a good idea?

Answer: It's a terrible idea.

The question and answers in this dataset were gathered using regular expressions from transcripts of five CNN series[3] containing interviews as well as the Switchboard corpus (Jurafsky et al., 1997). The focus of the original dataset was indirect answers to questions containing a gradable adjective. As discussed by (Westney, 1986) and others, it is possible to observe the phenomenon with a non-gradable scalar adjective, as in "Is the bathwater lukewarm already?".

The response may either omit the adjective (example 3.3), modify the adjective with an adverb (example 3.4) or negation (example 3.5), or use another adjective in it's place (example 3.2). Of the 224 question-answer pairs, 55.8% of them involved a response with another adjective. As this is the only situation where knowledge of scales is useful, we will focus on this type of response for the remainder of the discussion.

(3.3) Question: Is that a huge gap in the system?

Answer: It is a gap.

(3.4) Question: Were they happy?

Answer: They were so happy.

(3.5) Question: Are you bitter?

Answer: I'm not bitter because I'm a soldier.

[3]Available at `www.cnn.com/TRANSCRIPTS/(acd|ldt|le|lkl|sitroom).html`

Adjective	Frequency
good	25
right	7
confident	4
correct	4
big	3
nice	3
optimistic	3
acceptable	2
true	2
accurate	2

Table 3.6: Most frequent adjectives found in questions in IQAP dataset

Question Adjective	Answer Adjective	Frequency
good	*great*	5
correct	*true*	3
good	*excellent*	3
right	*correct*	2
good	*terrific*	2
lot	*few*	2
hard	*difficult*	2
nice	*beautiful*	2
right	*wrong*	2
good	*positive*	2

Table 3.7: Most frequent adjectives combinations found in IQAP dataset

Analyzing the IQAP dataset, we see that many of the same adjectives appear in questions, with only 69 unique adjectives appearing in the 125 questions. The most frequent adjectives found in the questions are given in table 3.6. Similarly, when an adjective is repeated in the dataset, the answer is often the same as well. The most frequent question answer combinations are found in table 3.7.

This is not a critique of the methodology used by de Marneffe, et al., but rather a reflection of the reality of data availability. This type of response has been studied

primarily in casual settings, of which there are fewer corpora available. Stivers and Hayashi term this type of answer as a transformative answer, and it is a method to resist the normal constraints of a yes/no question (2010). In their dataset, Stivers and Hayashi find that 35% of transformative answers are an attempt to correct a perceived issue with how the question was phrased. The strongest level of resisting the conventions of the question is to replace a word in the question with another. While it is not commented on how many questions are responded to this way, it is clear this is a natural and important phenomenon.

3.3.1 Constructing a Diverse IQAP

To build a more diverse set of indirect question pairs we chose to use transcripts from daytime American soap operas, similar to the corpus built by Davies (2011). Using the Stanford Parser we processed the corpus, selecting any yes-no questions that contained an adjective and whose response contained a different adjective and not any words like *yes* or *no*. While this returned a few hundred results, most were of the variety "Are you okay?". Having not found the diversity we were looking for, we next applied the same technique to dialogue found on the website Wikiquotes[4].

Using quotes from books, movies, television shows, comic books, and video games, we again found little diversity. The type of questions that elicit transformative or indirect answers are present, but the avoidance found in natural dialogue does not seem to be commonly use by writers.

To build a more diverse dataset we only gather the questions from the Wik-

[4]www.wikiquotes.org, April 1st 2017 databse dump

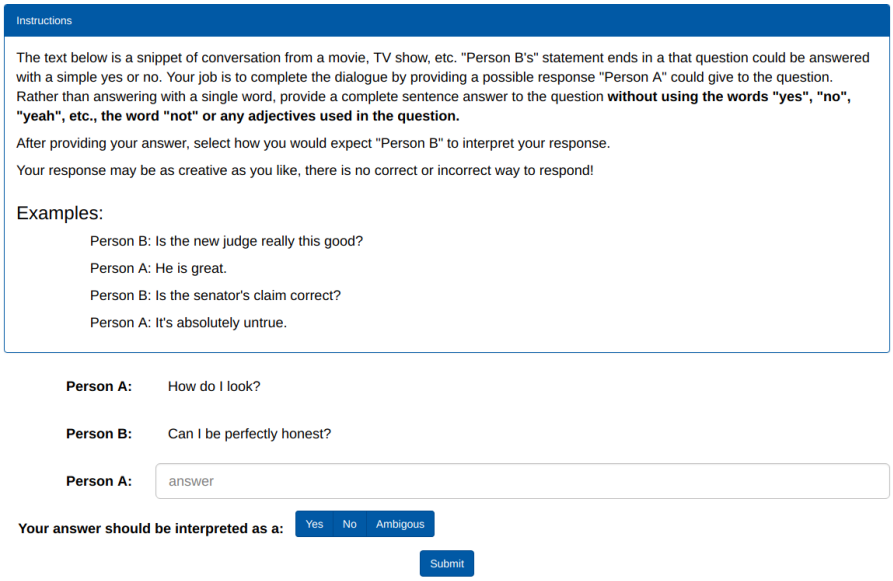

Figure 3.8: Interface used to gather indirect responses.

iquotes data, and ask informants on Amazon Mechanical Turk to respond to the question in an indirect way. The interface used for this task can be seen in figure 3.8. Originally questions like the one shown in example 3.6, where the adjective is not the focus of the question, were being returned from the corpus.

(3.6) Feel like a big boy, telling your big boy lies?

To remove this type of question, we add the constraint that questions must have a specific syntactic structure, namely that the adjective in the question must be an adjectival complement of the root verb of the question[5]. This constraint also removes some valid questions, but it produces a much higher quality dataset. This results in 347 questions, with each adjective only allowed to occur in three questions, and a total of 170 unique adjectives.

[5]As determined by using the spacy.io dependency parser

Each question was responded to by three informants. The informants' responses were shown along with the question to an additional 5 informants, who selected how the response should be interpreted, as a *yes* or a *no*.

3.3.2 Results

The informants provided 1041 answers, of which 255 were immediately discarded due to the informant being unsure if their responses should be interpreted as a *yes* or *no*.

Of the remaining 786 answers, 407 did not contain any adjectives. The prompt purposefully did not indicate that the answer should contain an adjectives, so we could investigate the commonness of responding with an adjective, although the examples presented in the instruction did contain an adjective. An example of this is shown in sentences 3.7 and 3.8.

(3.7) You're not very *bright* are you?

(3.8) I graduated in the top of my class.

Following these filtering steps, 44 answers which contained *yes*, *no*, *yeah*, or *not* were removed, along with 10 responses that answered the question with another question.

Finally, 89 responses containing obscene words were removed due to a desire to present the task on Mechanical Turk without an obscenity warning.

This left 237 question-answer pairs of the original 1041. The remaining questions used 124 unique adjectives, while the remaining answers used 171 unique adjectives.

221 combinations of adjectives in the questions and answers existed.

During manual inspection of the data we discovered that while many of the question-answer pairs did involve adjectives from the same scale, some were either alternatives but not scalar, such as sentences 3.9 and 3.10, some were not related in anyway, and in some the pair could be considered to constitute an ad-hoc scale.

(3.9) Round is round, am I right?

(3.10) If a square is square.

The 237 question-answer pairs were then presented to 5 additional informants, who provided their interpretation of the answer. Four informants' responses were removed prior to any further processing due to incorrectly answering both control questions.

Following this, 24 question-answer pairs were removed from the dataset due to a majority of respondents either indicating that they could not tell if the answer should be interpreted as a *yes* or *no*, or they believed that the question itself was not a yes-no question.

A further 14 question-answer pairs were removed because no majority opinion emerged about how the answer should be interpreted. Questions where the dominant interpretation by informant's was different than what the source of the answer indicated were also thrown out. After this pruning, several pairs remained that did not reflect a scalar alternative, even though the correct response was inferable by the respondents. Examples of this are shown sentences 3.11-3.14. Both of these pairs were interpreted to mean *yes* by all 5 informants, yet the adjectives used in them

Question-Answer Pair	Answer 3grams	Adj-Replaced 3grams	PMI
Q: Are you aware of the fact they've had shark attacks here? A: The last attack was six years ago.	the, last, attack	the, aware, attack	0
Q: Are you new? A: This is my first week	my, first, week	my, new, week	0.261
Q: Isn't he fantastic? A: I think he's a great person	a, great, person	a, fantastic, person	4.649

Table 3.8: Examples question-answer pairs demonstrating pruning process. Pairs with PMI less than 3 are pruned.

are *full*, *frustrated*,*ready* and *aware* and *last*, respectively. None of these adjective pairs occur on the same scale.

(3.11) Are you full of rage?

(3.12) I feel frustrated and ready to explode.

(3.13) Are you aware of the fact that they've had shark attacks here?

(3.14) The last attack was six years ago.

In order to remove these pairs, we calculate the PMI of the trigrams of each adjective in the answer, replacing the adjective with the one from the question. If a question or answer had more than on adjective, the average PMI was used. Example trigrams and their replacements are shown in table 3.8. Pairs with an average PMI of less than 3 were pruned. This does remove some legitimate pairs where the direct substitutability of the question adjective into the answer adjective isn't attested in the corpus, but a large majority of the removed pairs are non-scalar adjective pairs.

Following this pruning, 124 question-answer pairs are left. The final dataset contains 92 unique adjectives in the questions, with the most common, *big*, occurring

5 times. There are 107 unique adjectives in the answers, the most frequent being *good*, which occurs in 6 answers. There are 153 unique adjective combinations, due to some questions and answers having more than one adjective. The most frequent adjective combinations are *beautiful-lovely* and *big-huge*, which each occur 3 times.

3.4 Summary

In this chapter we presented the two datasets built for this dissertation. Our gold standard of scalar adjectives is the first gold standard created to not only order adjectives, but also consider which adjectives should be considered members of a scale. Our MIQAP dataset expands the variety of adjective combinations found question-answer pairs for use in evaluating adjective scales.

Chapter 4: Identifying Scale Members

In this chapter we introduce our methodology to identify other members of a scale given two adjectives that are members of that scale, which we refer to as seed words. To accomplish this, we combine a number of previous NLP approaches, including pattern based methods as well as lexical substitution. We discuss the various hyperparameters in the method by their impant on performance as well as the specific values that are best. Following this we analyze in detail specific instances of change to gain insight on why the particular hyperparameters are important and possible improvements that could be made. Our system achieves state of the art results on the gold standard dataset described in chapter 3.

4.1 Methodology

Our methodology consists of three phases. In the first phase, we use patterns from Sheinman & Tokunaga (2013) as well as de Melo & Bansal (2013) to gather exemplar sentences from a corpus, which are then used as input to a lexical substitution system. This produces a list of potential replacements for each seed word in each sentence, which are then aggregated using Cultural Consensus Theory, treating each sentence as a different informant. We explain this in more detail in the following

subsections, using the running example of initializing the process with the input of *hard* and *easy*.

4.1.1 Pattern-Based Exemplars

The patterns used in both Sheinman (2013) and De Melo (2013) take the form of "x or very y", in which x and y are replaced with seed words, producing a search string like "*hard* or very *easy*". Table 2.2 shows many other patterns that were used. We initialize these patterns with both seed words in each position, as we are not concerned with the semantic ordering of the words, but whether they are scalemates. For our example of *easy* and *hard*, the search strings of *easy or very hard* and *hard or very easy* will be used to find exemplar sentences. Searching the corpora returns exemplar sentences such as

(4.15) Books on numbers tend to be very *hard* or very *easy*.

(4.16) There are subtle differences , which will make for an *easy* or very *hard* installation, so the best advice is to check , check and then check again .

While this produces more sentences than if we insisted on order, the results can still be sparse depending on the seed words. To augment these we introduce a parameter called **expansion**, which when true, finds exemplar sentences using batches of seed words. We find exemplar sentences using a seed pair as before, and then substitute other seed pairs in the batch into the sentences with some probability, defined in initial experiments as .2. While these sentences may be felicitous or not, we hypothesize that the additional value of more datapoints will outweigh the less

exact semantic context. In particular, both the BalAdd and nPIC lexical substitution methods used in this work use context as defined by dependency parses, and so while the sentence as a unit appears incorrect, the dependencies may be much more plausible.

If *easy* and *hard* were in a batch with the seed words *cold-hot* and *tiny-minuscule*, substitution into sentence 4.15 would result in the sentences below. Substitution is based on word order in the seed pair, and no effort is made to place the word in the more likely position. Sentence 4.18 appears to be at least odd, but without this expansion step, the seed pair of *minuscule-tiny* would not produce any exemplar sentences, as they occur zero times in our corpus with any pattern. The corresponding dependency based contexts are shown in table 4.1, demonstrating the acceptability of the dependencies for substituted words.

(4.17) Books on numbers tend to be very *cold* or very *hot*.

(4.18) Books on numbers tend to be very *tiny* or very *minuscule*.

Sentence	Word	Dependencies
4.15	*hard*	(advmod^{-1}, *very*) (cc^{-1}, *or*) (conj^{-1}, *easy*) (acomp, *be*)
4.15	*easy*	(advmod^{-1}, *very*) (conj, *hard*)
4.17	*cold*	(advmod^{-1}, *very*) (cc^{-1}, *or*) (conj^{-1}, *hot*) (acomp, *be*)
4.17	*hot*	(advmod^{-1}, *very*) (conj, *cold*)
4.18	*tiny*	(advmod^{-1}, *very*) (cc^{-1}, *or*) (conj^{-1}, *minuscule*) (acomp, *be*)
4.18	*minuscule*	(advmod^{-1}, *very*) (conj, *tiny*)

Table 4.1: Dependencies of the original and substituted words in sentence 4.15. The dependencies appear much more plausible than the sentence as a whole. $^{-1}$ represents an inverse dependency, that is the second word listed is dependent on the word being discussed

4.1.2 Lexical Substitution

The sentences found as described above are used as input to a lexical substitution system. Specifically, for each sentence we provide the system with two instances, one where we set the target word as the first word and one where we set the target word as the second word. In prior work there have been two versions of the lexical substitution task investigated. The most common is when a system is provided a small list of candidates, usually obtained from annotators, as possible substitutions for a target word. Alternatively, no candidates may be provided and all words in the vocabulary are considered potential substitutes.

To save on computation while not over-specifying the problem, we set the possible candidates for substitution to all the adjectives in the 5000 most frequent words as calculated from the Corpus of Contemporary American English, excluding the appropriate seed adjectives[1] (Davies, 2008). This gives each target word 839 candidate substitutions.

Each sentence/target pair is processed with one of context2vec, BalAdd and nPIC, as described in Chapter 2, to produce a list of possible substitutes. The number of candidates returned is the second hyperparameter in our methodology, and values tested were the 2, 3, 5, 8, and 10.

We also examine a setting where the number of suggested candidates returned is a random value between 3 and 10. This is to simulate an informant listing a different number of substitutions for each sentence. The replacements for the indicated seed

[1]Available from http://www.wordfrequency.info/free.asp

Algo.	Sentence (4.15) *hard*	Sentence (4.15) *easy*	Sentence (4.16) *hard*	Sentence (4.16) *easy*	Sentence (4.17) *hot*	Sentence (4.17) *cold*
C2V	simple	simple	difficult	expensive	thin	mild
	fast	tough	quick	instant	weak	dark
	short	difficult	fast	unusual	wet	dry
	boring	boring	expensive	interesting	heavy	wet
	tough	automatic	smooth	obvious	dry	heavy
nPIC	tough	difficult	difficult	quick	wet	warm
	difficult	quick	tough	simple	warm	dry
	expensive	fast	fast	difficult	dry	mild
	tight	simple	soft	automatic	sunny	wet
	soft	tough	tight	fast	cool	cool
BalAdd	difficult	difficult	difficult	difficult	wet	wet
	tough	simple	tough	simple	warm	warm
	impossible	quick	impossible	quick	dry	dry
	soft	tough	nice	impossible	cool	cool
	harsh	fast	fast	neat	mild	sunny

Table 4.2: First 5 suggested substitutes for sentences (4.15) - (4.17) using the three algorithms studied.

words of sentences 4.15-4.17 are shown in table 4.2.

4.1.3 Cultural Consensus Theory

Pioneered by Romney et al., Cultural Consensus Theory (CCT) provides a way to weight an informant's response by their competency (Romney et al., 1987, 1986; Batchelder and Romney, 1989). In this application we take each sentence-target pair to be an informant, and weight the list of adjectives suggested by the substitution method accordingly. This allows us to aggregate the substitution candidates from the lexical substitution together into a scale. The application of CCT produces a list of words, which we take to be scalemates. Completing our example, with inputs *easy* and *hard*, one output of our system is the set {*tough, difficult*}, while the gold

standard is {*simple, difficult*}.

Weller points out CCT is especially advantageous over frequency aggregation when the options are close, giving the example of 48% answering yes and 52% answering no for a binary question (2007). More importantly for us, CCT also produces an easily interpretable score for each response, given in equation 3.1, where $G_k > 0$ indicates a response should be taken as part of the cultural consensus.

4.2 Experimental Setup

To examine the performance of our method, as well as the effects of expansion, the groups of seed words used for expansion, the lexical substitution measure, and number of candidates returned, we use the dataset described in chapter 3. We randomly chose 9 scales as a development set and put aside 3 for a test set. For each scale, we generate all possible pairs of words in that scale to use as input to our methodology. We evaluate using precision, recall, and F-score using the remaining words on the scale as the gold standard to test against.

The use of the expanded setting of our data creation mechanism means that the combinations between the seed word pairs in a batch is a factor. Testing all combinations would require running the methodology on 914,457,600 combinations of seed pairs. While this is not currently possible, to investigate the interaction between seed pairs as well as the general effect of seed words, we run 500 trials. The seed words for each trial were selected by using a 9-dimensional quasi-random generator, in order to cover as much of the space as possible (Sobol', 1967). This produces

batches containing each seed pair at least once and 88.77% of 2-way combinations between seed pairs. An example input for one trial is given in 4.19.

(4.19) (*small, large*), (*damp, wet*), (*smart, intelligent*), (*terrible, bad*), (*quick, speedy*), *(simple, difficult)*, (*several, many*), (*dark, light*), (*cold, hot*)

For each trial we gathered the exemplar sentences both with and without expansion, and then ran the 3 different lexical substitution measures. For each of these 6 combinations we returned a list of adjectives of 6 different lengths. Thus for each trial, we run 36 different combination of the pipeline for 9 scales.

The exemplar sentences in this work were gathered from the combination of UkWaC and WaCkyPedia (Baroni et al., 2009). Contex2vec is the version distributed with the original paper, which was trained on UkWaC, while the other methods are used as implemented by Roller (2016).

4.3 Hpyerparameter Anyalsis

We analyzed the data using within-subjects ANOVA, using the seed words as the subject identifiers. The results of ANOVA run against the F-scores are presented in table 4.3. The results for ANOVA over Precision and Recall are similar. The bold lines in the table indicate significant parameters. The expansion (E), length (L), and algorithm (A), as well as two and three way combinations of them, are the most significant. Due to the large number of instances, it is not surprising that many of the parameters were deemed significant by the ANOVA, although it is promising that the difference between trials given the same seed word was not found to be

significant.

	F value	p	η_p^2
Expanded (E)	**20657.152**	**0.00**	**.113**
Length (L)	**3021.619**	**0.00**	**.036**
Algorithm (A)	**1185.945**	**0.00**	**.035**
Trial (T)	0.048	0.827	0.000
E:L	**93.241**	**0.00**	**.003**
E:A	**1048.036**	**0.00**	**.013**
L:A	**21.666**	**0.00**	**.001**
E:T	1.164	0.281	0.000
L:T	0.023	1.000	0.000
A:T	0.303	0.739	0.000
E:L:A	**11.019**	**0.00**	**0.000**
E:L:T	0.037	0.999	0.000
E:A:T	0.290	0.9748	0.000
L:A:T	0.129	0.999	0.000
E:L:A:T	0.141	0.999	0.000

Table 4.3: ANOVA and η_p^2 with F-score as the dependent variable. Lines in bold represent significant parameters

To further investigate the effects of each parameter, we looked at its effect size, as measured by partial η^2 (denoted η_p^2), also shown in table 4.3. This shows that while the interaction between parameters has a significant effect, it accounts for very little of the variation seen in the data. The choice of whether to use the expanded dataset or not affects the precision, recall, and F score the most. The Tukey Post-Hoc test shows that using the expanded dataset increases F score by about 0.12, recall by 0.28, and precision by 0.09.

We analyze a reduced dataset consisting only of instances where the expansion was done, which reveals the algorithm is the next most important parameter, having an η_p^2 of 0.089 compared to an η_p^2 of 0.058 for the length parameters. The Tukey Post-Hoc test shows that there is a significant difference between the means of the

F-scores when using context2vec and BalAdd or nPIC, but not between nPIC and BalAdd. Because there is no significant difference, we remove only instances run using context2vec and analyze the data once again.

Having removed the largest cause of variation in the lexical substitution algorithm parameter, it is not surprising that the number of responses returned by the algorithm now has a larger effect than the algorithm itself. Almost all values of this parameter differ significantly from each other, with the exception of returning 2 or 5 words, which has no significant difference. Returning 3 words has the best F score, while returning 2 words produces significantly higher precision, and returning 10 words has a significantly higher recall. After this analysis, the configuration with the highest F score was using nPIC while returning 2 words per sentence-target pair, which achieves an F-score of 0.36522, slightly higher than returning 3 words while still using nPIC, which has an F-score of 0.36502.

Following this initial analysis, we further investigated the length and expansion parameters. To further investigate the length parameter, we tested returning all words which had a probability greater than 0.01, while using the expanded set of sentences and the nPIC and BalAdd substitution algorithms. We initially tried to set the length proportional to these probabilities, but the results were very poor and further analysis was not conducted.

Analyzing this data, is is found that unlike previous values for the number of words returned, using a value of returning words with a score greater than 0.01 produces significant differences between nPIC and BalAdd. Using 0.01 with BalAdd produces an F-score of 0.227, the lowest seen of any configuration using either BalAdd

or nPIC with the expanded setting, while using a value of 0.01 with nPIC produces an F-score of 0.377, the highest seen out of the same group. For this reason, we will evaluate the significance of the differences using only nPIC.

Among instances using nPIC and the expanded setting, 0.01 performs better than all other options on F-score, although not significantly better than using lengths of 2 or 3. The precision when using 0.01 sees a small but significant increase over the previous best configuration of 3, and an insignificant increase in recall over using 3 as well.

To further examine expansion, we introduced a third setting, proportional expansion, where the probability of a pair being used to create an artificial sentence is proportional to the number of sentences found using that pair and the number of sentences found using the most frequent pair. This choice was driven by early observations that blanket expansion involving seed words that already had many exemplar sentences tended to have a negative effect. We hoped to determine if tying the expansion probability to the relative number of sentences returned by a seed pair in a batch would reduce this.

For example, given any sentence, regardless of the pair used to find it provided it is not the pair being considered, the probability that a pair, e.g., *easy-hard*, will be substituted into that sentence is proportional to 1 minus the number of sentences already found using *easy-hard* over the highest number of sentences found using any pair in that batch.

This was evaluated using both nPIC and BalAdd lexical substitution methods and length settings of greater than 0.01, fixed length of 5 and 3, and a random length,

to cover various combinations of precision and recall. In all situations, this produces a small, but significant ($p < 0.001$) improvement in the mean recall, precision, and F score, over the purely random expansion used in prior experiments. Focusing on using nPIC, which is the most successful configuration, we see the difference between using length 3 and chosing all words with a score over 0.01 is not significant in regards to F or Recall, but using 0.01 produces significantly better precision. This simple improvement to the expansion step suggests that even further improvements can be made through more research on this one parameter alone.

Finally, we test the use of CCT more akin to a true ensemble method, providing it with the output for both the best performing configuration of nPIC (length = 0.01, proportional expansion) and the best performing configuration for BalAdd (length = 3, proportional expansion) and allowing it to aggregate over all the answers for each seed pair in each batch. Doing this increases the F score even further, and performs significantly better than either configuration on its own.

4.4 Evaluation

We denote the optimal configuration of parameters as **P0.1N** (Proportional expansion, words with probability > 0.01, nPIC), and determine another set of parameters **No10C** (No expansion, return 10 words, context2vec), by making the opposite decisions at each point. We denote the ensemble method as **ensemble**. Our results averaged over all trials and scales for each seed and then averaged over all seed means is presented in tables 4.4 and 4.5, along with a comparison to the best

configuration of the method described by van Miltenburg (denoted VM in tables 4.4 and 4.5), which uses the Moby Thesaurus to filter the list of pairs. The results in table 4.5 were computed by running all combinations of test scale seed pairs using the **P0.1N** configuration.

Precision and recall are calculated with respect to the remaining words in the gold standard scale, other than the two seed words.

Because van Miltenburg's method originally only returns pairs of words, we form scales out of them by taking the intersection of all words each seed word is in a pair with. These statistics are averaged over all possible seed pairs. We do not compare against any of the various clustering methods as these methods require the number of scales to be known and make the assumption that all words being clustered belong to one of these scales.

We do compare against using an embedding space and finding the words closest to the mean vector of the the two seed words. Numerous embedding spaces were examined, and we present the highest scoring one below, the freely available word2vec embeddings trained on the Google News corpus (Mikolov et al., 2013). Special caution should be taken with this comparison, as it was done under advantageous conditions of always selecting K closest words to the mean vector, where K was the number of remaining words on the gold standard scale. The other methods did not have access to this information.

The results for the test set on the best performing configurations was lower than simply using an ensemble based method, primarily due to poor precision. There are a number of potential reasons for this, and given that the word embedding

	P	R	F
Ensemble	0.496 (±0.08)	**0.414 (±0.05)**	**0.411 (±0.05)**
P0.1N	**0.566 (±0.09)**	0.389 (±0.06)	0.393 (±0.05)
word2vec	0.368 (±0.00)	0.368 (±0.00)	0.368 (±0.00)
No10C	0.067 (±0.002)	0.198 (±0.006)	0.095 (±0.002)
VM	0.188 (±0.398)	0.070 (±0.13)	0.090 (±0.16)

Table 4.4: Results of our **P0.1N** and **No10C** configurations compared to (van Miltenburg, 2015) and a simple word embedding model on the development set

	P	R	F
Ensemble+	**0.168 (±0.07)**	**0.275 (±0.12)**	**0.202 (±0.09)**
P0.1N+	0.130 (±0.07)	0.241 (±0.14)	0.159 (±0.09)
Ensemble	0.135 (±0.08)	**0.221 (±0.14)**	0.162 (±0.10)
P0.1N	0.105 (±0.07)	0.194 (±0.14)	0.130 (±0.08)
word2vec	**0.166 (±0.00)**	0.166 (±0.00)	**0.166 (±0.00)**
VM	0.090 (±0.25)	0.045 (±0.13)	0.056 (±0.15)

Table 4.5: Results of our **P0.1N** configuration compared to (van Miltenburg, 2015) on the test set

performs significantly lower on the test set than it did on the development set, we conclude that the 3 random chosen scales for the test set, {*ancient*,*old*,*fresh*,*new*}, {*same*,*alike*,*similar*,*different*} and {*hideous*,*ugly*,*pretty*,*beautiful*, *gorgeous*} are inherently more difficult, although whether this is due to a poorly constructed scales or other reasons is not evaluated in this document.

One observation we were able to make was that of the 360 possible combinations of seed pairs from these 3 scales, 54 batches return no sentences. This is because all three of the seed pairs in those batches return 0 sentences and thus there are no sentences to expand from. The seed pairs that return zero sentences are *hideous-ugly*, *ugly-gorgeous*, *beautiful-gorgeous*, *hideous-pretty*, *hideous-beautiful*, *hideous-gorgeous*, *same-alike*,*alike-similar*,*alike-different*, *ancient-fresh*,*old-fresh*, and *ancient-new*. To

overcome this sparsity we added two random seed words from two different scales from the development set to each batch, so that these batches are more likely to return sentences. This is denoted by **Ensemble+** and **P0.1N+** in table 4.5. By adding random seed pairs, only 13 of 360 batches fail to return sentences, yet the majority of the improvement is seen in pairs that already returned results rather than those that primarily occur in batches with 0 sentences.

As shown, with our **P0.1N** and **ensemble** configurations we achieve state-of-the-art on precision, recall, and F-score. We hypothesize that the good performance on recall over the pattern based method comes from our method's ability to produce words from patterns that are not attested in a corpus by using distributional semantics in the form of lexical substitution, while the increase on precision over traditional embeddings is due to using the words in context. In the next section we perform a detailed analysis of each of the parameters and discuss how each one helps and hurts in certain situations.

4.5 Discussion

In this section we look at the specific instances that see the most improvement and largest decrease in performance as measured by the F-score, given each parameter. Section 4.3 discusses which parameters have the largest effect, and in this section we seek to answer why these parameters cause those effects.

4.5.1 Expansion

The use of expansion unsurprisingly has the largest effect on seed pairs that would normally return no results from a corpus. In the original setting, the largest difference between no expansion and random expansion produces several instances of an increase of 1.0 across many trials, substitution methods, and candidate list lengths. As they failed to return any sentences originally, their precision and recall were both 0 and the F-score was undefined. One example is the pair *freezing* and *hot*. This large increase was achieved after the seed words were substituted into 10 sentences, 5 of them found with the seed pair *few/many*, 4 of them from the seed pair *simple/easy* and one from the seed pair *small/large*. Of course other factors play a role in this large increase, such as the fact that the length for this particular example was set to 2, the exact number of remaining words being elicited, but the large number of instances with an improvement of 1.0 in F-score demonstrates the value of expansion.

Conversely, using expansion hurts instances where those seed pairs already perform well. The most dramatic example of this in our data is using the seed pairs of *freezing* and *cold*, which see a reduction in F-score of 1. This instance occurs numerous times using the BalAdd lexical substitution method and the number of candidates returned set to 2. These words appear in 3 sentences in the corpora using the patterns, which increases to 7 sentences when using expansion in the particular instance we examine here. All additional sentences in this particular trial were originally found using the seed pair of *simple* and *easy*. Even though this

results in reasonably felicitous sentences, as shown in excerpts below, it is enough to return some unrelated words. Sentence fragment 4.21 proposes *frozen* and *wet* as substitutions for *freezing* and *hot*, and *wet* as substitutions for *cold*.

(4.20) very *simple* , very *easy* to navigate .

(4.21) very *freezing* , very *cold* to navigate .

Averaging over all lexical substitution methods and numbers of candidates returned, 26 seed pairs see a decrease in F score when using the expanded sentences, while 83 pairs see an increase. The average decrease is 0.025, and the average increase is 0.216.

When we analyze the expansion setting using proportional generation of extra sentences as compared to no expansion, we again see several instances of increases of 1.0, occurring in similar circumstances as above. The largest decrease in F score was 0.80, found when of using *easy* and *simple* as seed words. In this particular trial, the number of sentences the pair appeared in was originally 15, and after expansion, the pair appeared in 36 sentences. This is due to the precedence of *moist* and *wet* in this batch, which returns 42 sentences. This suggests that while adding sentences proportionally to the most frequent seed pair in the batch is a better approximation, it can still produce negative results in a few situations. Possible solutions to this would be a maximum number of sentences after which expansion is never performed, or determining to expand proportionally based on the average number of sentences returned for each pair in the batch, or some other summary statistic. Further research is needed to determine the optimal ways in which expansion occurs and how to

achieve this optimum more consistently.

When we average over all lengths investigated with both no expansion and proportional expansion, we find that 94 of 111 pairs see an increase using proportional expansion. The average increase in F score is 0.26. The 14 pairs which see a decrease have an average change of -0.01, while 3 pairs see no change.

Comparing to the random expansion used initially, 92 seed pairs see an increase when using proportional expansion, 15 see a decrease, and 4 see no change. The average increase is 0.02 while the average decrease is 0.007.

4.5.2 Length

As we did in the previous section, we will look at a case where changing only the length improves the F-score and one where it decreases it. Because there are several different values for the length parameters, we will be looking at the largest and smallest differences between maximum and minimum F scores while holding all hyperparameters but the length constant. The largest improvement of an F-score comes when using the seed pair *slow* and *speedy*.

Using a length of 2 achieves an F score of 1.0, while using a length of 8 produces an undefined F score due to having both a precision and recall of 0. When comparing these two cases, neither the sentences themselves or the algorithm changed, only the number of suggested substitutions per sentence. This example highlights the nuances behind using CCT, as using a length of 10 has an extremely high recall, just as using a length of 2, while a length of 8 has both a recall and precision of 0. The

reason for this is that in positions 6, 7, and 8, many words are returned for multiple sentences, including *dry*, *sunny*, and *easy*. This causes sentences producing those words to be viewed as better informants. Their answers then count more and the correct answers are overwhelmed by them. When the length is increased to 10, the correlation drops between these informants, allowing the correct answers to at least make it into the final result, along with many incorrect answers.

The minimum change observed for a trial where the F scores were not all already undefined was between a length of 2 and a length 5 when using the seed pair *tiny* and *enormous*. The F score across all length settings ranged from 0.4 to 0.4444. This small variation is due to relatively balanced fluctuations in precision is recall, where a length of 2 has a precision of 0.677, but because the scale in this case has 6 remaining members, it would be very rare to get all 6 with such a short length. A length of 8 sacrifices precision to achieve a .33 recall.

Length is the trickiest parameter to optimize in our opinion, and requires a great deal of further research. The attraction of the method presented here is that CCT includes a built in cut off value, but length clearly influences the number of words on the scale produced by CCT. As shown in the additional experiments, basing the number of words on the probability assigned by the lexical substation method shows promise, at least when using the nPIC measure, but we are confident that even more accurate results could be obtained through a combination of more work on the length parameter along with improved substitution metrics that more accurately reflect the likelihood of substitution.

4.5.3 Trial

The last source of variation we examine is the influence the other seed words have on the performance of a system when expanding the number of sentences. To examine this, we looked at the maximum and minimum F-scores for each seed, using the proportional expansion parameter. We will look at the pairs with the largest and smallest distance in more detail, excluding those that see no change because they were always the most frequent pair in their batch, and thus never had any additional sentences created for them. The most stable seed pair across all trials is *bad* and *awesome*, when using a length of 5 and BalAdd. This pair appears in 18 trials, and has an F score of 0.7272 across all trials. Without expansion, this pair returns 0 sentences, and using expansion returns a mean of 32.72 sentences. The most common seed pairs *bad* and *awesome* occur with are given in table 4.6.

Returning to the particular trials involved, the final output of the system in terms of the words returned is always the same, although the order does vary. The words are *awful*, *terrible*, *horrible*, *good*, and *nasty*. For comparison, the 5 words returned as closest to the mean of those two words embedded using word2vec are *good*, *horrible*, *terrible*, *amazing*, and *unbelievable*.

At the other end of the spectrum are two trials involving the pair *freezing* and *warm*. One trial has a precision and recall of 0, having returned *dry*, *frozen*, and *wet* as the scale members. This seed pair returns no sentences naturally from the corpora, and in the trial where it performs worst, is in 6 sentences. Five of the sentences originally occurred with *easy* and *hard*, while one sentence was found using

the pair *few-several*. The highest scoring trial of this pair has an F-score of 1.0, in which the seed words appear in 21 total sentences, most originating with the seed pair *hard-difficult* or *few-several*. All of this suggests that further refinements to the expansion process are needed.

Seed pair	Freq.
simple, *easy*	5
some, *several*	5
cold, *warm*	5
freezing, *cold*	4
light, *bright*	4
simple, *hard*	4
stupid, *intelligent*	4
several, *many*	4
slow, *quick*	4
slow, *speedy*	4

Table 4.6: Frequency of seed pairs found in batches with *bad,awesome*

4.6 Summary

In this chapter we have presented our methodology and analyzed many parameter settings which influence our methodology. We have shown that generating artificial data through string replacement in a loosely guided fashion significantly improves the results. We have also shown that CCT shows promise as a general ensemble method for aggregating ranked lists. Finally, we analyzed each parameter for an in depth understanding of the values and the difference they make, as a guide towards further work.

Chapter 5: Ordering Scale Members

In this chapter we present our solution for ordering the words contained in a scale. Our method is an improvement to de Melo and Bansal's Mixed Integer Linear Programming (MILP) approach to ordering of words (2013). DeMelo and Bansal's methodology is itself an improvement of the strictly pattern based method of Sheinman and Tokunaga (2009). By reasoning over an entire half[1] of a scale, de Melo and Bansal are able to overcome some of the sparsity typically encountered when using pattern based methodology. Even with this improvement, the MILP system remains under-defined in many cases. By using lexical substitution, we are able to generate reasonably correct sentences that can then be interpreted using traditional pattern based methods. Applying this step, we see an improvement in scale order.

Our methodology includes several other improvements. While both Sheinman & Tokunaga and DeMelo & Bansal order subscales, we take an entire scale as input. We do still split the scales into two subscales for part of the ordering procedure, but this is done automatically using clustering. Furthermore we introduce several new patterns used to determine which side of the scale should be positive.

[1]Half here refers to a positive half and negative half and should not suggest an even numerical divide. We prefer the use of subscale instead.

5.1 Methodology

de Melo and Bansal's insight was that by assigning each pair of words a score indicating how likely the first word of the pair is to be to the left of (less than) the second word in the pair, and combining these scores, the sparseness of pattern based methods can be overcome. Unfortunately there are still many situations where either not enough information is collected, or a specific word participates in no patterns with any other word on the scale. This is not surprising, as while language can and often does contain many redundancies, by using a particular word on the scale, the speaker is communicating that intensity directly, and there is no need to reference another word on the scale.

To motivate the need for this, we will use a running example of the scale ⟨*minuscule*, *tiny*, *small*, *big*, *large*, *huge*, *enormous*, *gigantic*⟩.

5.1.1 Partitioning the Scale

We believe it is important to address and order all the words on the scale together as many scalar adjectives indicate a value inherently less than a word traditionally thought to be opposite of it. For example *small* indicates an object having less size than the same object described as *big*. That being said, it is rare to find words on opposite sides of the scale used in the kind of constructions that indicate order. The phrase "*small but not big*" strikes us as odd, yet "*small but not minuscule*" is acceptable and attested in corpora. For this reason we automatically cluster the scale into two halves and attempt to reduce the sparsity in each half,

rather than over the scale as a whole. In early work we found that most logical errors in sentences produced by lexical substitution were made when attempting to create a sentence containing a word from each half, for example *tiny* and *gigantic*.

To partition the scale we used K-means (k=2) clustering on paragram-phrase word vectors (Wieting et al., 2016). The paragram embeddings result from learning sentence embeddings that minimize the cosine distance between two sentences in vector space that are listed as paraphrases in the XL version of the Paraphrase Database (Ganitkevitch et al., 2013), while maximizing the difference between each sentence and the pair and several negative examples picked from the same corpus. As a composition function, the authors found that simple averaging of the word vectors produced the best results, but it is important to point out that this average, along with other composition functions investigated, are used during training and not just as a post processing step. The word vectors produced this way achieve state-of-the-art performance on many semantic textual similarity (STS) testsets, but also are high quality general word vectors for use in a wide variety of tasks.

5.1.2 Augmentation

As with DeMelo and Bansal, we first search a corpus, in this word ukWaC and Wackypedia (Baroni et al., 2009) , with several patterns. The patterns used by DeMelo and Bansal differ slightly from those used by Sheinman and Tokunaga, and are displayed in table 2.3. Patterns are classified into weak-strong patterns, where the first word is less than the second word, and strong-weak patterns, where the

opposite is true.

For each word-pair combination in a subscale, we search the corpus and return all sentences matching the patterns. We then send all sentences to a lexical substitution algorithm. For this application we focus on context2vec, which preliminary studies indicated works best, perhaps because it more accurately captures the semantics needed for ordering words. Then we randomly sample the sentences proportional to the lexical substitution scores, and substitute accordingly.

In some instances, a subscale will return no sentences for all possible pairs of words. In this case, we relax the query to the corpora, searching for all sentences that contain one of the words in the patterns, with the other slot filled with a wild card. Sentences returned this way are still sent to a lexical substitution system, but now we are searching for the best replacement for the word in the slot where the wildcard was, among the remaining words on the scale. These substituted sentences along with the attested sentences from the corpus are then used in DeMelo and Bansal's methodology to assign scores between words.

There are several hyperparameters that control this process to produce slightly different sentences.

5.1.2.1 Normalization

For each word in the subscale that is not in the pair, we normalize the scores from context2vec in one of two ways. One way is to normalize each column, that is normalize the word over sentences to find the most likely sentence that word can be

substituted into. For example, if sentences were found using the pattern "*small or even tiny*", the scores for *minuscule* being substituted for *small* would be normalized against each other and the scores for *minuscule* being a substitute for *tiny* would be normalized against each other.

The other setting for this hyperparameter is to normalize the scores for each sentence against each other. For sentences found using the pattern "*not huge but still big*", the scores of the remaining three words on the scale would be normalized against each other for each sentence independently.

5.1.2.2 Independent or Combined Normalization

In either setting for normalization, the scores can be normalized for each slot independently, or with relation to the other slot in the sentence. In the independent setting, using sentence 5.22 as an example, the score of *gigantic* being substituted for *large* and the score of *gigantic* being substituted for *huge* have no impact on each other during the normalization. In the combined setting, the opposite is true, so that *gigantic* is much more likely to be substituted for *huge* in a particular sentence than *large*.

(5.22) This is a Victorian, gabled house, large but not huge.

5.1.2.3 Sampling Proportionally

The final hyperparameter we investigate during the generation phase is weighting the likelihood of substitution by the number of sentences already seen between

that word and the word it would be in a pair with if the substitution was made. This can be used to ensure that pairs with a large amount of evidence are not frequently found in artificial data, while pairs with little or no existing sentences will be more likely to be sampled.

5.1.3 Ordering

The DeMelo and Bansal algorithm was originally intended for intensity only, which means that one of the subscales needs to be reversed. For example, in the original methodology, *minuscule* is more intense than *small* and thus is to the right if a line is imagined. In our application, it is important that *minuscule* is less than *small*. This is achieved simply by negating the scores of all words on the left subscale. To determine the left half scale, we introduce several new patterns. These are based on binomial constructions and other constructions that while presenting no semantic reasoning for the words to be in a certain order, they are often found that way. The patterns used are shown in table 5.1.

x ones y ones
too x or too y
x and y
x or y

Table 5.1: Patterns used to orient the two subscales relative to each other. x is the word in the right subscale.

Using these new patterns we define a new metric, M, that is used to estimate how likely a word is to be to the left of a word in the other subscale. The formula for M as well as how it is used in the revised score function are shown in equations

5.1 - 5.4. To determine the ordering of the two subscales, we sum all the M scores of the cross product of the two subscales, and if the score is positive, the first word grouping is placed on the left and if the score is negative the second grouping is taken to be the left subscale.

$$P_3 = \sum_{p_3 \in P_{\text{middle}}} \text{count}(p_3) \tag{5.1}$$

$$M_1 = \frac{1}{P_3} \sum_{p_3 \in P_{\text{middle}}} \text{count}(p_3(a_1, a_2)) \tag{5.2}$$

$$M_2 = \frac{1}{P_3} \sum_{p_3 \in P_{\text{middle}}} \text{count}(p_3(a_2, a_1)) \tag{5.3}$$

$$score(a_1, a_2) = \begin{cases} \frac{(W_1 - S_1) - (W_2 - S_2)}{\text{count}(a_1) \cdot \text{count}(a_2)} & \text{if } a_1 \text{ and } a_2 \text{ in same subscale} \\ \\ \frac{M_2 - M_1}{\text{count}(a_1) \cdot \text{count}(a_2)} & \text{otherwise} \end{cases} \tag{5.4}$$

There is one hyperparameter we investigate in regards to the new metric M. The naive application of this metric would weight the evidence between *tiny* and *gigantic* as equally as *big* and *small*. We believe that the patterns shown in table 5.1 are more robust the less extreme the words involved in them are. To weight the evidence of this accordingly, we introduce a penalized version of the score function, which weights the relation by how extreme the words are. To determine this, the two subscales are ordered independently to determine the intensities, and then this information is used as weights. This is shown in equation 5.6.

$$penalty = \frac{1}{(\textbf{index of}(a_1) + 1) \cdot (\textbf{index of}(a_2) + 1)} \tag{5.5}$$

$$score(a_1, a_2) = \begin{cases} \frac{(W_1 - S_1) - (W_2 - S_2)}{\text{count}(a_1) \cdot \text{count}(a_2)} & \text{if } a_1 \text{ and } a_2 \text{ in same subscale} \\ \frac{M_2 - M_1}{\text{count}(a_1) \cdot \text{count}(a_2)} penalty & \text{otherwise} \end{cases} \tag{5.6}$$

We run the MILP using the objective function that uses this information, but also do a version where the sides are ordered completely independently and then simply placed on the appropriate side of the scale, which we call the naive method.

5.1.4 New Model

One additional change we investigate is a constraint on the MILP defined in equation 2.8, that prevents the x values of a word, that is their estimated position on the scale, from being the same. The complete definition of the MILP with the new constraint is given in equation 5.7. It is logical to allow words, for example *big* and *large*, to occupy the same space on the scale if an outside resource like WordNet is used, as was done in the original implementation. We do not use any external resources in our implementation and we hypothesize that allowing ties causes many pairs of words with insufficient evidence to be assigned the same value on the scale, whether correct or not. To prevent this, we add the following constraints to the definition of the MILP, which have the effect of equally spacing all words on the scale.

$$
\begin{aligned}
\text{maximize} \quad & \textstyle\sum_{(i,j)\notin E}(w_{ij} - s_{ij}) \cdot score(a_i, a_j) - \sum_{(i,j)\in E}(w_{ij} + s_{ij})C \\
\text{subject to} \quad & \\
& d_{ij} = x_j - x_i && \forall i, j \in \{1, \ldots, N\} \\
& d_{ij} - w_{ij}C \leq 0 && \forall i, j \in \{1, \ldots, N\} \\
& d_{ij} + (1 - w_{ij})C > 0 && \forall i, j \in \{1, \ldots, N\} \\
& d_{ij} + s_{ij}C \geq 0 && \forall i, j \in \{1, \ldots, N\} \\
& d_{ij} - (1 - s_{ij})C < 0 && \forall i, j \in \{1, \ldots, N\} \\
& x_i \in [0, 1] && \forall i \in \{1, \ldots, N\} \\
& w_{ij} \in \{0, 1\} && \forall i, j \in \{1, \ldots, N\} \\
& s_{ij} \in \{0, 1\} && \forall i, j \in \{1, \ldots, N\} \\
& \mathbf{d_{ij}} = \mathbf{u_{ij}} - \mathbf{v_{ij}} && \forall i, j \in \{1, \ldots, N\} \\
& \mathbf{u_{ij}} \leq \mathbf{1} - \mathbf{y_{ij}} && \forall i, j \in \{1, \ldots, N\} \\
& \mathbf{v_{ij}} \leq \mathbf{y_{ij}} && \forall i, j \in \{1, \ldots, N\} \\
& \mathbf{u_{ij}} + \mathbf{v_{ij}} \geq \tfrac{\mathbf{1}}{\mathbf{N+1}} && \forall i, j \in \{1, \ldots, N\}
\end{aligned}
\tag{5.7}
$$

5.2 Experimental Design

To compare the effectiveness of our proposed changes, we split the gold standard data into the same development and test sets used in the previous chapter. For each of the 9 scales, we run 100 trials of generating augmented data for each hyperparameter combination for a total of 800 trials for each scale. For each trial both the original and more restricted version of the MILP are run over the data a total of 20 times,

due to some instability of the linear solver based on the order of it's arguments. This gives us 2000 trials for each setting for each scale. In addition we repeat the same experiment for the modified score function that penalizes the M metric, although this only affects running the MILP over the entire system and not running it over a single subscale or the naive method.

As baselines, we run the same hyperparemters as before, but replace the lexical substitution system with either random selection or use cosine similarity between a candidate substitute and its target as calculated using word2vec embeddings. Using word2vec emeddings allows us to investigate the importance of context to this methodology.

5.3 Results

Much like the previous chapter, we examine the optimal hyperparamter settings. We again use within-subjects ANOVA, using the scale and type of ordering together (subscale, naively constructed full scale, and full scale) as the subject identifier. The results of ANOVA and the effect size as measured by η_p^2 using ρ as the dependent variable are shown in table 5.2.

With regards to ρ, using the augmented dataset or not has the largest effect size, with Tukey's Post-Hoc test reporting that using the augmented data results in an estimated increase of 0.134 over using just truly attested data. Eliminating trials using the unaugmented data and running ANOVA again, the next largest effect size is whether the normalization is done independently for each slot or in a combined

	F value	p	η_p^2
Trial	0.526	0.4684	1.14e-07
Iteration	0.048	0.8260	1.05e-08
Augmentation	**9496.827**	**0.0000**	**4.21e-03**
Penalty	**115.226**	**0.0000**	**2.50e-05**
Combo	**46.264**	**0.0000**	**1.00e-05**
Balanced	**35.237**	**0.0000**	**7.65e-06**
Normalized	**11.119**	**0.0008**	**2.41e-06**
MILP Model	**1026.764**	**0.0000**	**2.23e-04**

Table 5.2: Results of ANOVA using ρ as the dependent variable. Bold lines are statistically significant

manner. The difference between the means however was less significant than the difference between using a balancing weight to discourage substitutions when a large amount of evidence exists, which Tukey's Post-Hoc test estimates improves ρ by 0.036. The next largest effect size was whether the M score was penalized or not, and although this only affects using the MILP with the full scale at once, the analysis suggests using the penalized version of the M score to produce better results. Finally, when considering the remaining trials, the largest effect is the interaction between the type of normalization used and if combination is used or not. The more effective hyperparameter setting is to use column-wise normalization and normalizing each word slot independently. There is no significant difference between using the original MILP definition or the one that prohibits ties, but using the definition proposed in this work does produce slightly higher ρs.

The best set of hyperparameters when considering ρ are then: using the augmented dataset, using the balancing weight, using the penalized M score, normalizing the lexical substitution scores independently and column-wise, and the MILP defini-

	Subscale	Full-Scale	Naive
This Work	0.444	0.543	0.498
Random Baseline	0.443	0.595	0.518
W2V Baseline	**0.471**	**0.600**	**0.521**
DMB original paper	0.347	0.415	0.509
DMB original paper + New MILP Def	0.330	0.354	0.469

Table 5.3: Performance in terms of ρ of ordering methods

tion that prohibits ties. The average ρs produced by this setting compared to the best hyperparameter settings for the baselines and original method are shown in table 5.3.

When we perform ANOVA with pairwise accuracy as the dependent variable, the hyperparameter with the largest effect is whether the MILP definition prohibits ties or not. This is not surprising, as our problem definition does not allow ties, our definition of pairwise accuracy does not count ties as correct, and the new MILP definition does not allow ties. Following this same logic through, the best hyperparameters in order of effect size are weighting the chance of substitution by the number of sentences with that pair already present, using the penalty in the M-score, and column-wise and independent interaction. This is the same set of hyperparameters as with regards to ρ, but in a different order of importance. The comparison between this method and baselines is shown in table 5.4.

While the results above improve upon previous work, it is interesting that picking a word randomly performs so well. The highest performing word2vec instance also reduce to random picking, as all use the column normalization, and as each candidate will always have the same score to replace a given word, this reduces to

	Subscale	Full-Scale	Naive
This Work	0.727	0.746	0.736
Random Baseline	0.713	0.768	0.735
W2V Baseline	**0.730**	**0.770**	**0.740**
DMB original paper	0.402	0.621	0.599
DMB original paper + New MILP Def	0.670	0.673	0.715

Table 5.4: Performance in terms of Pairwise Accuracy of ordering methods

Normalization	Weighted	Ind. Normalization	This Work	Random	word2vec
Column-Wise	Yes	No	**0.567**	0.511	0.507
Column-Wise	Yes	Yes	**0.584**	0.499	0.478
Column-Wise	No	No	**0.607**	0.532	0.553
Column-Wise	No	Yes	**0.599**	0.539	0.542
Row-Wise	Yes	No	**0.510**	0.474	0.476
Row-Wise	Yes	Yes	**0.529**	0.467	0.483
Row-Wise	No	No	**0.533**	0.500	0.511
Row-Wise	Yes	Yes	**0.535**	0.494	0.507

Table 5.5: Percent of new sentences generated that demonstrate the correct relationship

a uniform distribution. There are two possible reasons for this. One is that we are picking from a limited set of choices, and sometimes any choice results in a logically correct phrase. As an example, we will look at sentence 5.23 below. If *huge* is replaced with any other word from the scale randomly, the sentence is still correct. Of course this is an extreme case, but even in a case containing the phrase *huge but not enormous*, two-thirds of random substitutions into *huge* will be correct, while one-third of substitutions for *enormous* will be correct. When looking at the percentage of correct sentences, we see that using a lexical substitution algorithm does produce more logically consistent sentences, but this improvement is not enough that it makes a difference in the MILP. The percent of generated sentences that are correct given each hyperparameter are shown in table 5.5.

(5.23) yes , the freemasons has reopened , and i know i don't get out much , but i think the interior and the garden are a very big , if not huge , improvement

The second reason is that we are averaging over all subscales, and while it is logical for some hyperparmaters to be better for some scales than others, we believe it is instructive to look at the performance of different parameters with regard the structure of the scale. The structure of the scale in terms of how many words are on it is observable by the system and therefore can be used to make deterministic decisions without knowledge of the words themselves or their correct ordering. For this discussion we consider 4 different possibilities:

I. The subscale contains 3 words, and at least one pair of words returns sentences

II. The subscale contains 4 or more words, and at least one pair of words returns

sentences

III. The subscale contains 3 or more words, and no sentences are found using any of the pairs

IV. The Subscale contains 2 words, and no sentences are found between them

One scenario we will not discuss is when the subscale contains two words and at least one sentence is found using those words. In this case, augmentation is never applied, so it is irrelevant for this analysis.

For all scale structures, the hyperparameter with the largest effect is the normalization direction. Scale structures I, II, and IV perform best when using column-wise normalization, while scale structure III performs best when using row-wise normalization. This deviation makes sense, as scales with structure III have no attested sentences in the corpus, and thus the sentences being substituted into carry no guarantee that the words are being used in a scalar manner. By sampling with respect to how likely a word is to be used in a sentence compared to the other possibilities, we are more likely to produce a logical sentence.

Scale structure III is also the only variation to perform significantly better when normalizing the candidate scores together rather than independently. To the best of our knowledge, the is the result of an anomaly with the random sampling, as normalizing the candidate scores together only has an effect when there is more than one slot to fill. Scale structure II also performs better when normalizing the candidate scores together, but the difference is not significant.

Scale structure I is unique in that it performs better when there is no weighting

Scale Structure	Normalization	Weighted by Existing Pairs	Ind. Normalization
I	Column-Wise	No	Yes
II	Column-Wise	Yes	No
III	Row-Wise	Yes	No
IV	Column-Wise	Yes	Yes

Table 5.6: Best Configuration given Scale Structure

	I	II	III	IV	Average using best
This Work	**0.875**	0.532	**0.542**	**0.391**	0.486
Random Baseline	0.838	0.787	0.466	0.156	0.501
W2V Baseline	0.849	**0.788**	0.536	0.290	**0.511**
(De Melo and Bansal, 2013)	0.866	0.504	*	*	0.346
dM&B + New MILP Def.	0.758	0.521	-0.046	0.028	0.335

Table 5.7: Best ρ given scale structure
Best ρ given scale structure. * Represents value cannot be determined because linear solver cannot be run

done based on the pairs already attested.

The best configurations for each scale structure type are shown in table 5.6. The ρs for each scale structure given its best configuration are shown in table 5.7 while the pairwise accuracy is shown in table 5.8. The configurations were held constant for the hyperparameters in this work, while the baselines were always selected to be their highest. Even though the random baseline achieves a higher ρ score in many situations, those same configurations are not competitive when looking at pairwise accuracy, and vice-versa.

Following this analysis, it is clear that there is more to be looked into behind why the random baselines perform so well on scale structure II. One possibility is that the increased number of sentences, and thus the reduced number of words pairs

	I	II	III	IV	Average using best
This Work	**0.917**	0.726	**0.764**	**0.695**	**0.747**
Random Baseline	0.866	**0.826**	0.691	0.578	0.741
W2V Baseline	0.866	0.825	0.722	0.645	**0.747**
(De Melo and Bansal, 2013)	0.667	0.502	*	*	0.402
dM&B + New MILP Def.	0.839	0.752	0.485	0.514	0.673

Table 5.8: Pairwise accuracy given scale structure. Represents value cannot be determined because linear solver cannot be run

	% Correct	Density	% New Sentences
Group I	15.57	1.48	-24.98
Group II	-0.36	-8.37	-41.74
Group III	3.99	-3.36	NA
Group IV	7.05	0.00	NA

Table 5.9: Difference in correctness, density, and percentage of new sentences between lexical substitution and random substitution for each type of scale structure.

with no information under the random substitution leads to increased performance. The average difference between using lexical substitution and random substitution is shown in table 5.9 for the percentage of sentences that are logically consistent, the density of the system, and the number of new sentences generated as a percentage of the number of original sentences.

	% Correct	Density	% New Sentences
Group I	0.148818	0.017778	-0.254560
Group II	-0.024875	-0.071042	-0.379694
Group III	0.036533	-0.040034	NA
Group IV	0.064046	0.00	NA

Table 5.10: Difference in correctness, density, and percentage of new sentences between lexical substitution and word2vec context-free substitution for each type of scale structure.

From this we can see that the graphs for ordering scales with structure II not only are less dense when using lexical substitution, they are also slightly less correct, although we would not expect this small difference in correctness to lead to such a large difference in ordering. When we look at the actual scores generated using the patterns and sentences, we find that lexical substitution is only slightly less correct.

5.4 Application to Data from (De Melo and Bansal, 2013)

In their original paper, de Melo and Bansal tested their method on 88 subscales. These scales were all collected from WordNet, but were pruned by human annotators to remove words that did not belong on the given scale. Unlike our set up, words in this dataset were allowed to have the same intensity. Because this dataset consists exclusively of subscales, we modify our methodology to remove the clustering prior to both augmentation and ordering.

The original paper used the Google N-grams corpus (Brants and Franz, 2006), which contained approximately 95 billion sentences and 1 trillion words. In this work, our combined corpora of Wackypeadia and UkWaC contains about 3 billion words. They also introduce a version that uses WordNet to explicitly move items in the same synset together. We do not compare against this version as one goal of this work is to produce adjective scales without relying on existing resources.

Because this dataset contains ties, we ran our methodology with the optimal configuration, but ran it using both MILP definitions, to see if the ties impacted the performance of one over the other. The results are shown in table 5.11. We report

MILP Def.	Dataset	Pairwise Acc.	τ	$\lvert\tau\rvert$	ρ	$\lvert\rho\rvert$
New	Augmented	0.614	0.340	0.517	0.393	0.610
New	Unuagmented	0.601	0.314	0.505	0.349	0.593
New	Random	**0.637**	0.392	0.530	0.447	0.618
Old	Augmented	0.578	0.416	**0.562**	0.460	**0.625**
Old	Unaugmented	0.506	0.436	0.506	0.462	0.539
Old	Random	0.604	**0.445**	0.559	**0.491**	0.624
(De Melo and Bansal, 2013)		0.699	0.57	0.65	0.64	0.73

Table 5.11: Performance on dataset from (de Melo & Bansal 2013)

both τ and ρ as well as their absolute values as done in the original paper. As noted by de Melo and Bansal, the relative ordering of a subscale can be consistent between annotators, but the direction may be reversed.

The original implementation performs the best on τ and ρ, while using augmented data performs best on $|\tau|$ and $|\rho|$. This suggests that the additional data more correctly separates out the words on the scale, but has the potential to do it in the improper order. One solution to this is to use a separate process to determine the overall orientation of the scale, and only use the MILP to determine the relative ordering.

The only metric where the definition of the MILP that prohibits ties performs the best is pairwise accuracy. We believe this is due to the fact that prohibiting ties, in conjunction with the augmented data, forces words that have little evidence to one extreme or the other. For example, in the subscale $\langle$ *big, large, huge, enormous, gigantic*$\rangle$ if *gigantic* is forced to the far left, the ρ and τ values will suffer far more than the pairwise accuracy score.

For most applications, such as indirect question answer, a higher score on

pairwise accuracy is more beneficial as only two words are being compared. One instance where a higher ρ value may be desired is if the scales derived from this process were further analyzed to give numerical sentiment intensity scores which may then be used in sentiment analysis.

5.5 Summary

In this chapter we have shown how lexical substitution can be used to generate artificial data that is correct enough to improve the performance of a pattern based methodology. We have shown that the particular hyperparameters affect subscales differently based on their structure and the number of edges missing. We have also shown that a random substitution baseline performs well, surpassing the lexical substitution method in many instances. All of this suggests that artificial data generation is a promising avenue of future work for pattern based methods, and that more work is needed to determine how to do better than using random substitution.

Chapter 6: Learning and Ordering Scales from Noisy Seeds

In this chapter, we detail our efforts to build a large resource of adjective scales. We discuss the various design decisions needed to apply the methodologies discussed to a large, noisy input and present a human evaluation of the results. Finally we use the resource to complete the original IQAP and MIQAP tasks.

6.1 Determining Scale Membership

In chapter 4, we introduce our method to determine the members of a scale, given two words known to be on the scale. We now demonstrate the use of that same methodology in less ideal conditions. In this chapter we only use the highest performing configuration of hyperparameters: proportional expansion and an ensemble of nPIC returning words with a score higher than 0.01, and BalAdd returning the top 3 words.

6.1.1 Selecting Seed Words

For seed words, we use two different sources, which we will keep separate for analysis purposes. The first set of seed words are all pairs of adjectives listed as antonyms in WordNet, of which there are 1152. The second source are pairs found

according to van Miltenburg's pattern-based method for finding and ordering pairs of scalar adjectives. This set of seed words contains 1810 pairs. Both sets of seed words have been filtered to only include words that occur in the vocabulary of lexical substitution methods.

Both of these sources are inherently noisy. While WordNet is a manually constructed resource, words listed as antonymous are not guaranteed to be scalar. Examples range from borderline cases such as *broken* and *unbroken*, to clearly non-scalar adjectives such as *adoptive* and *biological*.

The scalar adjectives from Van Miltenburg's method alternatively are noisy in that they may not belong to the same property and either consist of a more ad-hoc scale, or simply are erroneous. Examples from this methodology include *valid* and *new* as well as *heavy* and *due*.

Batches were composed of 5 or 6 seed words each, with seed pairs chosen based on their frequency of occurring the in patterns used in constructing scales. For example, the first word in a batch for WordNet would be one of the 230 most frequent pairs, while the second word would be from next 230 words, although in reserve order. The first batch in WordNet consist of the most frequent pair, the 460th most frequent pair, the 461st most frequent pair, the 890th most frequent pair and the 891st most frequent pair.

6.1.2 Pruning Scales

With such noisy data, not all scales returned can be expected to be high quality data. Furthermore, unlike the scales used in chapter 4, the seed words may be members of a valid scale, but may be the only words on the scale and thus nothing else should be returned. To prune invalid scales, we rely on the eigenvalue ratio available to us as part of the output from cultural consensus theory. As a reminder, in anthropological research, the ratio between the first and second eigenvalues found during factor analysis indicates the level of consensus about a topic in a culture. The accepted value to indicate consensus in literature is 3.0 for use in human studies.

To our knowledge, this is the first work that uses CCT with computationally generated answers, and thus there is no accepted ratio. Therefore we stick with the accepted ratio of 3.0.

This leaves us with 668 scales from the WordNet seed words, and 216 scales from the VM seed words.

6.2 Human Evaluation of Membership

To evaluate the intrinsic quality of the dataset, we again turn to Mechanical Turk. Informants were presented with 10 groups of adjectives, with each group consisting of the words suggested by the method along with the two seed words. The informants were asked to indicate which words did not belong in each group. The words were shuffled and no indication was given as to which words were the seed words. The detailed instructions and interface can be seen in figure 6.1. Each

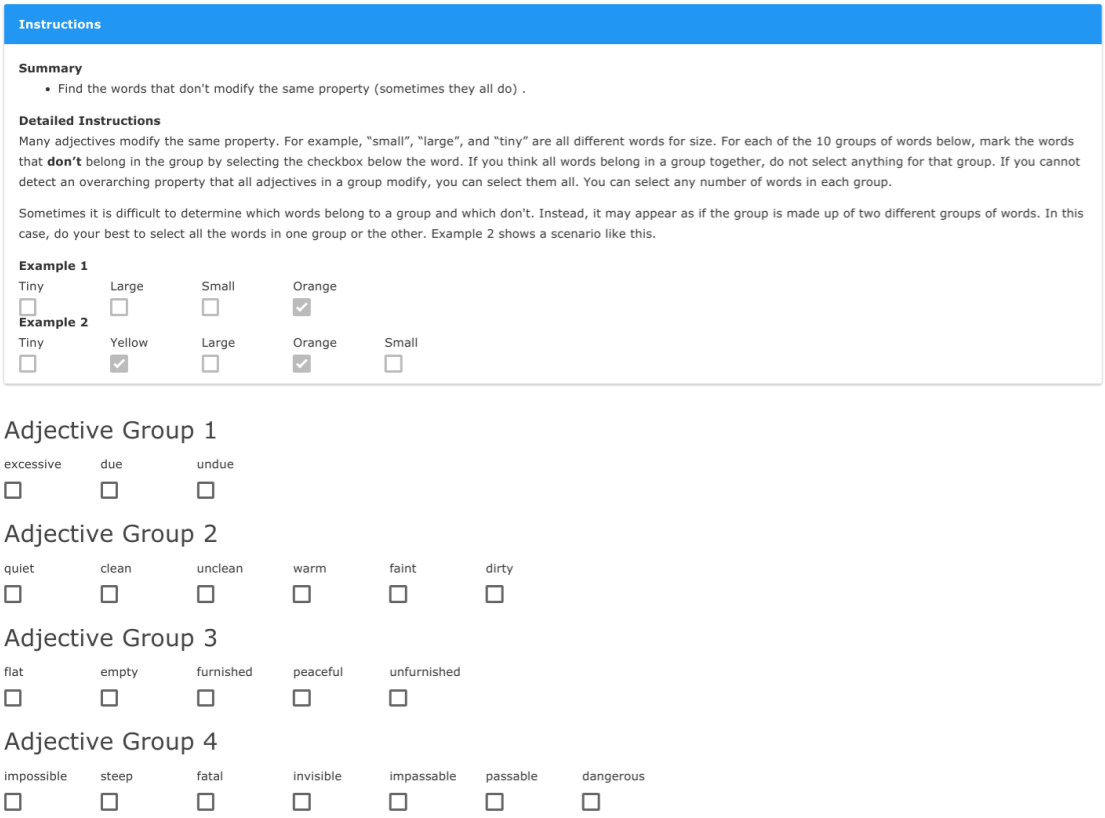

Figure 6.1: Interface used by informants to indicate which words do not belong to each scale.

potential scale was evaluated by 5 informants.

6.2.1 Scales from WordNet seeds

An average of 68.45% of words were determined to belong to each scale, when using the most conservative approach of removing a word if it had been selected as not belonging to the scale by at least one informant. Only 33 words of 3106 potential words were removed by all 5 informants, while 1628 words were never selected for removal. 22 scales had all members removed at least once, while 26 scales had all but one word removed at least once. 145 scales had no members removed. There

is an average of 4.65 words per scale after removing words selected by at least one informant. This increases to 4.97 words per scale if those scales that have 0 or 1 word remaining are not included in the calculation.

While these results look promising, we believe they should be interpreted with extreme caution. The task of selecting which words do not belong in a group can be difficult for even the most experienced of researchers, and informants may have had difficulty with the assignment. Some of these difficulties are easily identifiable, such as selecting to remove *distant* from the scale of {*close,closer,closest,distant*}, which we attribute to selecting the single word on the negative side of the scale, something we had seen previously in our analysis of the gold standard in chapter 3. Even when all the words on the scale should not be there, we notice the phenomenon of selecting either the one negatively or positively oriented word, such as removing *virtuous* from {*cruel, stupid, wicked, virtuous*}, where the author believes *stupid* should have been removed.

Another common situation we found was when the seed words, and thus usually the words suggested, were not scalar. We have no way of knowing this a priori, and were hoping that the informants would help identify these cases, but we believe the informants focused on selecting words that didn't belong in a general sense rather than ones that don't belong on the same scale.

Even more troublesome are the instances where no words were removed from scales that had either no or a very limited overarching property. One example from the dataset is the proposed scale of {*supportive, isolated, naked, false, empty, embarrassed, valid, hostile, unable, vulnerable*}, from which no words were ever

selected as not belonging.

Possible remedies for this issue are using a greater number of informants for each scale, selecting informants from a different pool other than Mechanical Turk, making the directions more explicit that the words should be scalar, and changing the interface to present less scales at once.

6.2.2 Scales from VM Seeds

An average of 60.61% of words were determined to belong to each scale by human annotators when using the seed words produced by Van Miltenburg's method. 23 of the 653 potential words were removed by all 5 informants, while 221 were never removed. 11 scales had all members removed at least once while 12 scales had all but one word removed at least once. 25 scales had no members removed.

From a qualitative point of view, it appears that the seed words from van Miltenburg's method produced better results in the sense that the informant had an easier time selecting the words that did and did not belong. Some of the same pitfalls as before were observed, such as informants selecting the single negative or positive member, and not checking if they were all members of the same scale, rather than just related, but we did observe less scales with zero members removed that were questionable.

One thing we did notice was that the 214 scales suggested using the VM seeds were not as unique. For example 6 of the 25 scales with no members removed contained the word *vital* and other words for IMPORTANCE. The average number

of words in a scale after removing the words checked by informants is 3.037. This increases to 3.338 words per scale if those scales that have 0 or 1 word remaining are removed.

A random sampling of scales using each set of seed words is shown in table 6.1, where bold words were seed words, and italicized words were checked by at least one informant as not belonging.

WordNet	Van Miltenburg
thick, toxic, dense, **compressible, incompressible**	fundamental, essential, vital, *crucial*, **important**, **key**
intelligent, tender, *symbolic*, vocal, rational, conscious, *naked*, biped, ***quadruped***	purple, pink, blue, yellow, ***full***, **red**
impossible, *necessary*, attractive, *essential*, *useful*, **desirable**, **undesirable**	*rare*, *far*, later, ***early***, **late**
used, dried, crowded, ***lined, unlined***	*useful*, *attractive*, suitable, *visible*, **accessible**, **available**
invisible, lonely, mild, toxic, gentle, aggressive, frequent, painful, faint, subtle, ***diurnal***, ***nocturnal***	specific, distinctive, unique, **exclusive**, **special**
cruel, mysterious, biological, *logical*, artificial, ***natural***, **unnatural**	extensive, *frequent*, *common*, **universal**, **widespread**
increased, extended, improved, **contracted**, **expanded**	unprecedented, distinctive, unusual, **rare**, **unique**
broken, dying, mixed, dirty, pale, purple, ***stained, unstained***	*obvious*, *necessary*, impossible, **inevitable**, **possible**
frozen, isolated, dried, empty, **drained**, **undrained**	*easy, complicated,* ***complex, simple***
damn, weird, ridiculous, scary, silly, **real**, **unreal**	fundamental, necessary, vital, crucial, ***central***, essential

Table 6.1: Sample scale memberships using two different seed sets. Bold words are seed words while italicized words were indicated as incorrect by a human annotator

Dataset	Avg % of words in WN	Avg % of words in FN	Avg % of words in VM seeds
VM Unfiltered	100%	77.04%	20.56%
VM Filtered	100%	76.89%	20.13%
WN Unfiltered	99.63%	58.81%	3.20 %
WN Filtered	99.60%	59.58%	3.14 %

Table 6.2: Percentage of new words discovered using membership identification methodology

6.3 Comparison of Membership with Existing Resources

In addition to asking human annotators to judge if words belong together, we compare the scales generated against two popular existing resources, WordNet and FrameNet. The first comparison we make is to look at how many words proposed as members of a scale already exist in the resources. We refer to the scales produced directly by the membership identification methodology as unfiltered and those from which words selected as not belonging by informants are removed as filtered. The summary statistics about the words' prevalence in other resources is shown in table 6.2.

Looking at this, we can see we aren't learning a particularly high number of words that aren't already covered by these resources. It is reassuring that we are learning scalar relationships between words that aren't learned through the coprora based method of van Miltengburg. The more interesting question though is how many items in a scale are contained in the same structure for a given resource. In a perfect world, every item on a scale would belong to the same dumbbell structure in WordNet, and the same frame in FrameNet. Table 6.3 shows the average number of

Dataset	Avg number of attributes in WN	Avg % of words belonging to attributes	Avg % of frames in FN	Avg % of words belonging to frames
VM Unfiltered	1.91	44.38%	2.43	77.97%
VM Filtered	1.26	41.62%	1.74	78.34%
WN Unfiltered	1.70	28.73%	2.76	49.74%
WN Filtered	1.31	28.26%	1.74	49.85%

Table 6.3: Average number of attributes and frames per scale.

attributes or frames found per scale. This was found by determining the attribute with the largest number of words shared, removing them, and then repeating until the attributes for all members of the scale were found, or there were no more attributes to be considered. Table 6.3 also shows the average percent of words per scale that belong to these attributes.

The low number of attributes/frames per scale suggests that the members being proposed are likely related. It is not unexpected that they are greater than one, as it is well known that certain groups of related adjectives occur as attributes of different nouns in WordNet. The number of words that aren't associated with any attribute or frame shows that this resource not only captures a new relationship between known words, but can be used as a way to expand the coverage of existing resources.

6.4 Determining Scale Order

After determining the scale membership as described above, we use the methodology introduced in chapter 5 to determine the order of the words on a scale.

We used both the direct output from the membership identification for WordNet and VM seeds, as well as the set of scales produced by removing all words selected at least once by informants as not belonging to the scale. Furthermore, scales that were proper subsets of another scale were removed along with any duplicate scales. This results in 665 scales found using the WordNet seeds, 602 scales when these were corrected by human annotators, 188 scales found using the VM seeds, and 119 scales when these were corrected by human annotators.

6.5 Human Evaluation of Scale Order

To gauge how well the scales were ordered without a gold standard to refer to, we constructed a task to post on Mechanical Turk. Informants were presented 10 scales at a time and asked if they agreed with the ordering presented, which was the result of running the MLIP over augmented data, or if they thought it was incorrect. If the informant selected that the ordering presented was incorrect, the words turned into interactive buttons that could be dragged to another location on the scale. An example of this interface is shown in figure 6.2. Each scale was evaluated by 5 informants.

We calculated the ρ value for each scale with respect to each informant. If an informant indicated they agreed with the order presented, we assigned a ρ value of 1.0 to this pair. Some informants selected they did not agree with the ordering, but provided no alternative ordering. For these instances, we assigned a ρ of 0. The average ρ's for the 4 sets of scales are shown in table 6.4.

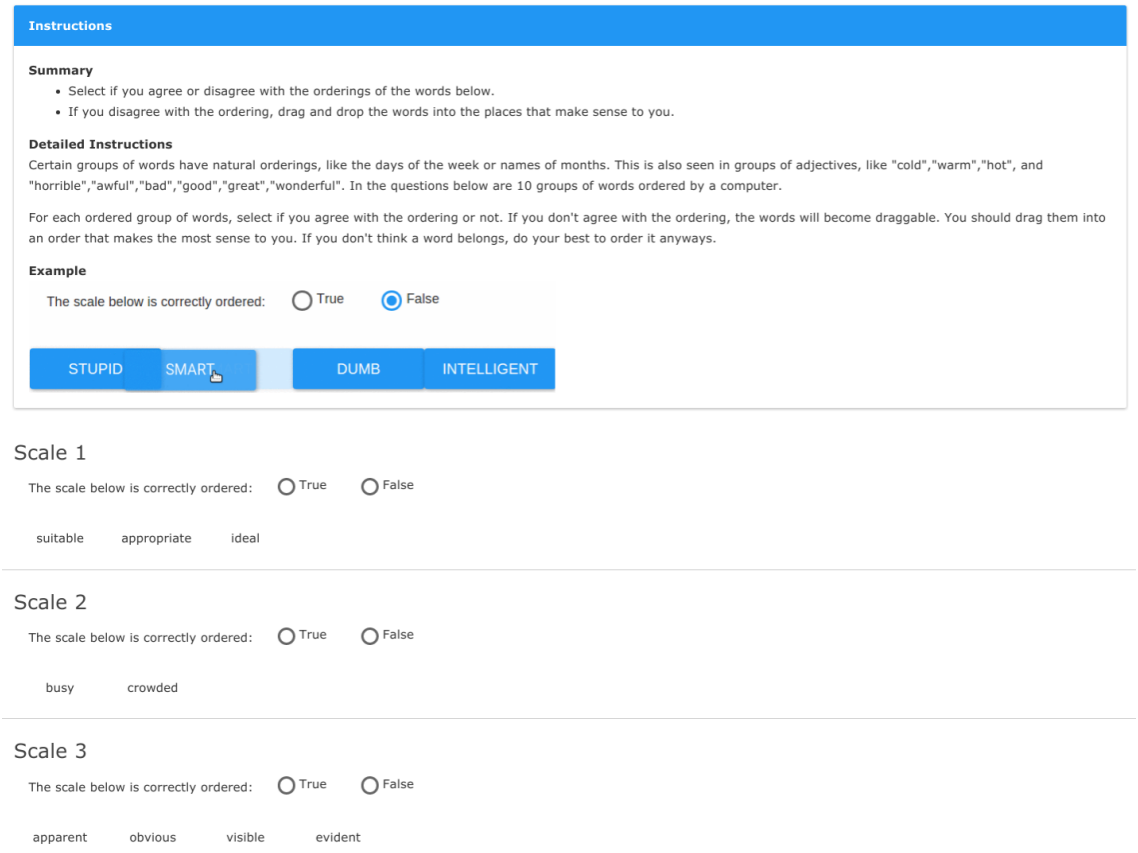

Figure 6.2: Interface used by informants to indicate which scales are ordered correct.

Dataset	ρ	IA ρ
VM Unfiltered	0.65	0.25
VM Filtered	0.73	0.26
WN Unfiltered	0.70	0.18
WN Filtered	0.72	0.21

Table 6.4: ρ for human evaluation of scales.

These results show that informants agreed with the orderings when the input had previously been filtered to remove extraneous words. This is intuitive, but further research is needed to determine if the better inputs lead to better output, or if the reduced length of scales in the filtered cases makes an informant more likely to accept the ordering. We also examined the average ρ between all informants that indicated the order was incorrect for each scale. These results are presented in table 6.4 as well. It is clear that annotators can agree that an ordering is wrong, but have difficulty indicated what the correct order should be. This may be an indication again that the input to the ordering methodology was in fact the issue, and not the ordering itself, as the ρ's seen here are higher than those seen in chapter 5, although in chapter 5 the gold standard order was created from a random initial ordering, not a suggested potential ordering as we did here.

6.6 Indirect Question Answer Pairs

Initially we intended to use the the scales learned above to solve the Indirect Question-Answer Pair (IQAP) task, but we found that many pairs in both datasets were either not on any scale, or not on a scale together. Because of this, we repeated the steps above, without the human filtering, to two sets of adjective pairs, one from IQAP and one from MIQAP.

If the adjective used in the answer was stronger, that is ranked more to the right than the adjective used in the question, we inferred that the answer should be interpreted as a *yes*. If an adjective-pair appeared on more than one scale, we

Testset	Methodology	Acc. (*yes*)	Acc. (*unconfirmed*)
IQAP	This Work	0.592	0.51
IQAP	(De Melo and Bansal, 2013)	0.632	0.55
MIQAP	This Work	0.601	0.403
MIQAP	(De Melo and Bansal, 2013)	0.573	0.371

Table 6.5: Accuracy on IQAP and MIQAP testsets.
Accuracy on IQAP and MIQAP testsets. Label in parentheses represents default value.

performed this inference using each scale, and then selected the more frequent answer. If there was more than one adjective in the question or answer, we again compared each pair of adjectives and took the more common answer, *yes* or *no*. In the IQAP set, if negation was indicated, we reverse the comparison, so that more *no*'s from the individual adjective pair comparisons would lead to a final inference of *yes*. We calculate the accuracy with two different default values to return if no scale is found, *yes* or *unconfirmed*.

The accuracy on each dataset is shown in table 6.5.

6.6.1 Prior Work

The IQAP testset was originally introduced in (de Marneffe et al., 2010) to investigate pragmatic phenomena. Using scores from movie reviews they determine an expected rating for each word. From this they determine if a word is more intense than another word. Using this data, they achieved a 60% accuracy on the IQAP testset. Since then numerous approaches using vector based representations have been proposed (Mohtarami et al., 2012, 2013; Kim and de Marneffe, 2013; Kim et al., 2016). These score significantly higher, up to 0.83 accuracy, but require an external

data source to identify the antonym of the adjective used in the answer.

While the difference between our scores and other work is partially a reflection on improvements that can be made with the methodology, it also reflects the difference between semantics and pragmatics. Indirect question answering is a pragmatic task that we approached with semantic data. While this is effective in many cases, there are some instances where a pure semantic interpretation leads to the wrong inference. Another issue we have noticed is that the same adjective pair may occur on multiple scales, and have differing orders on these scales.

This highlights the importance of the membership elicitation portion and show a need to investigate how close scales need to be before they are merged in some fashion.

6.7 Summary

In this chapter we have shown the application of our methodology to large and noisy input. We have shown that the membership identification relies on good input data. Regardless of the members found, the ordering methodology performs decently according to our human evaluation. The results on indirect question answering show that our method increases the number of questions that can be answered without having to guess a default value.

Chapter 7: Conclusion

It this dissertation we have produced a series of methods that, when used in conjunction, produce an ordered set of scalar adjectives, given two words on the scale. Using these we have produced and evaluated a resource of adjective scales. We have shown how lexical substitution provides a viable source of augmented data when the existing data is too sparse. We have also produced two new datasets: a gold standard of adjective scales for use in development of identification and ordering methods and a more diverse set of indirect question answer pairs for use as an evaluation.

7.1 Limitations

One of the largest limitations of this work is that development was carried out using such a small set of adjective scales. It is our hope that the scales generated here and the future work carried out on them provides a larger resource.

The other major limitation of this work is that it was only carried out in English, but we believe it would be especially beneficial with other languages that tend to lack the large scale resources available for English. While everything in this work should be language agnostic, testing must be carried out to prove this empirically. In order to apply this to another language, the following are required:

lexical substitution methods, training data for lexical substitution, patterns for seed extraction and membership elicitation, word vectors for clustering, and potentially a set of paraphrases to learn the word vectors.

7.2 Future Work

All the experiments and user studies done in this dissertation have many possible extensions. We will present them in the order they were presented in the dissertation.

With regards to the creation of gold scales, one extension is to elicit more scales, especially examples of words appearing on multiple scales. The saliency of the relationship in speakers' lexicons is another area of further study. In both the scale member elicitation in chapter 3 as well as the manual evaluation in chapter 6, we noticed that informants had trouble giving words for both sides of the scale. It is unclear if this reflects the actual structure of the lexicon, or simply a greater familiarity and ability to reflect about the antonymy and synonymy relationships rather than scalar relationship.

A possible solution to this is to use the visual metaphor of the scale, presenting the informant with three words already placed on the scale, either manually or through use of an automatic method, and allowing them to add as many other words along the scale as they wished. This would also combine the two elicitation steps into one, although we still suggest a second user study to gather a gold standard ordering from all words deemed to be salient rather than just the ones proposed by

an individual informant.

When constructing an indirect question answer dataset, a more thorough and linguistically sound method of filtering out questions should be developed rather than the rather crude method of using PMI. Something we would change in the elicitation of answers is not limiting the number of questions each adjective appears in before asking for responses, as many of the responses ended up being removed in the various pruning steps.

We see several areas of future work building on the methodology presented in chapter 4. One is a more detailed investigation into the ensemble set up of the task, using multiple sources, possibly even beyond lexical substitution, to generate the words provided to CCT. Also in relation to the use of CCT, the relationship between the eigenvalue given by CCT and human judgments of the effectiveness of the method should be studied.

We also see room for further work in how the extension phase of the methodology is carried out. One possible mechanism is to use language modeling to determine the probability of the sentence after the substitutions were made, or to use lexical substitution methods themselves to determine the probability of making both substitutions.

In chapter 5, we used a much smaller corpus than previous studies using mixed integer linear programming, but there is no reason both corpora cannot be used. The large Google n-grams corpus (Brants and Franz, 2006) cannot be used with lexical substitution as it does not provide enough context. A smaller corpus such could be used to produce the augmented sentences, while count from the the large

corpus be used directly in the MILP.

Extension to the definition of the MILP itself are an other area of future work. Ruppenhoffer demonstrated that the collocation of degree adverbs with adjectives is a promising technique for ordering adjectives, especially those of duration and size. This information could be incorporated as another source of ordering information.

Finally, we hypothesize that a reason that random substitution performs so well as a method for data augmentation is that it leads to more dense graphs than what can be found using lexical substitution. The relationship between graph density and performance on the ordering task is something that should be looked at more rigorously.

Following ordering the words in relation to each other, a more exacting relationship could be discovered based on the the value the words represent on the scale. This could be done partially through methods similar to Hickman, et al.(2015), but not all adjectives are found with numerical values in text. Another possibility is using the distance between two words in any number of embedding models to move the words to particular values while maintaining the already established overall order.

In chapter 6, it is clear that more work needs to be put into the phrasing of how informants are asked to chose which words don't belong on the scale. Potential options for this are making the scalar relationship more explicit in the instructions, or as suggested above with gold standard elicitation, present the words ordered on a scale.

Another future task suggested from our work in chapter 6 is to develop a classification method to determine if the words the MILP is being asked to order

consist of a whole scale or subscale. This is important because no clustering should be performed if the words only constitute a subscale. Finally, when inferring if the answer to a question is yes or no, we believe that it is important to have access to the entire scale, as one could easily answer a question involving the adjective *bad* with the words *horrible* or *great*. The inference changes if the words are on opposite sides of the midpoint of the scale, and thus it is important to find and store not just the order of the scale, but the midpoint as well.

Bibliography

Nabil Abdullah and Richard A Frost. 2005. Adjectives : A uniform semantic approach. In Advances in Artificial Intelligence: 18th Conference of the Canadian Society for Computational Studies of Intelligence. Springer, pages 330–41.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In COLING-ACL.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Lang Resources & Evaluation 43(3):209–226.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices : Representing adjective-noun constructions in semantic space. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pages 1183–1193.

William H Batchelder and A Kimball Romney. 1988. Test theory without an answer key. Psychometrika 53(1):71–92.

William H Batchelder and A Kimball Romney. 1989. New results in test theory without an answer key. In Mathematical Psychology in Progress, Springer, pages 229–248.

Sarah Benor and Roger Levy. 2006. The chicken or the egg? a probabilistic analysis of english binomials. Language 82(2):233–278.

Brent Berlin and Paul Kay. 1969. Basic Color Terms: Their Universality and Evolution. University of California Press.

Manfred Bierwisch. 1989. The semantics of gradation. In Dimensional Adjectives, Springer.

Dwight Bolinger. 1977. Neutrality, norm, and bias. Indiana University Linguistics Club, Bloomington, Ind.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram version 1. LDC2006T13. DVD. Linguistic Data Consortium.

Geoffrey L. Bursill-Hall. 1995. Linguistics in the later middle ages. In Concise History of the Language Sciences, Pergamon, pages 130 – 137.

Michael G. Carter. 1973. An arab grammarian of the eighth century a. d.: A contribution to the history of linguistics. Journal of the American Oriental Society 93(2):146–157.

Joseph B Casagrande and Kenneth L Hale. 1967. Semantic relationships in papago folk-definitions. In Studies in Southwestern ethnolinguistics, Mouton, pages 165–193.

Alan Cruse. 2011. Meaning in Language: An Introduction to Semantics and Pragmatics. Oxford University Press, USA.

Mark Davies. 2008. The Corpus of Contemporary American English (COCA): 520 million words, 1990-present. Available online at http://corpus.byu.edu/coca/.

Mark Davies. 2011. Corpus of American Soap Operas: 100 million words. Available online at http://corpus.byu.edu/soap/.

Marie-Catherine de Marneffe, Christopher D Manning, and Christopher Potts. 2010. Was it good? it was provocative. learning the meaning of scalar adjectives. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pages 167–176.

G De Melo and M Bansal. 2013. Good, great, excellent: Global inference of semantic intensities. Transactions of the Association for Computational Linguistics 1:279–290.

R Dixon. 1977. Where have all the adjectives gone? Studies in Language 1(1):19–80.

R. M. W. Dixon and Alexandra Y. Aikhenvald, editors. 2006. Adjective Classes: A Cross-Linguistic Typology. Oxford University Press, USA.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In NAACL. Association for Computational Linguistics.

Manaal Faruqui and Chris Dyer. 2015. Non-distributional word vector representations. In ACL. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA.

John R Firth. 1957. A synopsis of linguistic theory, 1930-1955. In Studies in linguistic analysis, Blackwell.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In Proceedings of NAACL-HLT. Association for Computational Linguistics, Atlanta, Georgia, pages 758–764.

Derek Gross, Ute Fischer, and George A Miller. 1989. The organization of adjectival meanings. Journal of Memory and Language 28(1):92–106.

Derek Gross and Katherine J Miller. 1990. Adjectives in wordnet. International Journal of lexicography 3(4):265–277.

Zellig S Harris. 1954. Distributional structure. Word 10(2-3):146–162.

Matthias Hartung and Anette Frank. 2010. A structured vector space model for hidden attribute meaning in adjective-noun phrases. In Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, pages 430–438.

Matthias Hartung and Anette Frank. 2011. Exploring supervised lda models for assigning attributes to adjective-noun phrases. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pages 540–551.

V Hatzivassiloglou and KR McKeown. 1993. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In Proceedings of the 31st annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, pages 172–182.

L Hickman, J Taylor, and V Raskin. 2015. Fuzzy lexical acquisition of adjectives. In Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC), 2015 Annual Conference of the North American. pages 1–6.

Fred W. Householder. 1995. Aristotle and the stoics on language. In Concise History of the Language Sciences, Pergamon, Amsterdam, pages 93 – 99.

Dan Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. Technical Report 97-02, University of Colorado, Boulder Institute of Cognitive Science.

Kyoko Kanzaki, Qing Ma, and Eiko Yamamoto. 2006. Acquiring concept hierarchies of adjectives. In Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead: 21st International Conference. Springer, pages 430–441.

C Kennedy and L McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. Language 81(2):345–381.

Christopher Kennedy. 1997. Projecting the adjective : The syntax and semantics of gradability and comparison. Ph.D. thesis, University of Calinfornia, Santa Cruz.

Christopher Kennedy. 2012. Adjectives. In Russell, G. and D. Graff Fara, editor, Routledge Companion to Philosophy of Language, Routledge., pages 328–341.

Joo-Kyung Kim and Marie-Catherine de Marneffe. 2013. Deriving adjectival scales from continuous space word representations. In EMNLP. Association for Computational Linguistics, pages 1625–1630.

Joo-Kyung Kim, Marie-Catherine de Marneffe, and Eric Fosler-Lussier. 2016. Adjusting word embeddings with semantic intensity orders. In ACL 2016 workshop on Representation Learning for NLP (RepL4NLP). Association for Computational Linguistics, pages 62–69.

Maria Koptjevskaja-Tamm and Ekaterina V. Rakhilina. 2006. "Some like it hot": On the semantics of temperature adjectives in Russian and Swedish. Language Typology and Universals 59(3):253–269.

A Lehrer and K Lehrer. 1982. Antonymy. Linguistics and philosophy 5(4):483–501.

Omer Levy and Yoav Goldberg. 2014a. Dependency-Based word embeddings. In ACL. Association for Computational Linguistics, pages 302–308.

Omer Levy and Yoav Goldberg. 2014b. Neural word embedding as implicit matrix factorization. In Advances in neural information processing systems. pages 2177–2185.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics 3:211–225.

Magnus Ljung. 1974. Some remarks on antonymy. Language 50(1):74–88.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In Proceedings of the 4th International Workshop on Semantic Evaluations. Association for Computational Linguistics, pages 48–53.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional lstm. In Proceedings of CONLL. Association for Computational Linguistics.

Oren Melamud, Omer Levy, and Ido Dagan. 2015. A simple word embedding model for lexical substitution. In Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. pages 1–7.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 .

Mitra Mohtarami, Hadi Amiri, Man Lan, Thanh Phu Tran, and Chew Lim Tan. 2012. Sense sentiment similarity: An analysis. In AAAI. Association for the Advancement of Artificial Intelligence.

Mitra Mohtarami, Man Lan, and Chew Lim Tan. 2013. From semantic to emotional space. In AAAI. Association for the Advancement of Artificial Intelligence.

Victoria Lynn Muehleisen. 1997. Antonymy and semantic range in English. Ph.D. thesis, Northwestern University.

Carita Paradis. 2001. Adjectives and boundedness. Cognitive Linguistics 12:247–271.

Victor Raskin and S. Nirenburg. 1995. Lexical semantics of adjectives. Technical Report MCCS-95-288, Computing Research Labroratory, New Mexico State University, Las Cruces, New Mexico.

Douglas Raybeck and Douglas Herrmann. 1996. Antonymy and semantic relations: The case for a linguistic universal. Cross-Cultural Research 30(2):154–183.

Stephen Roller and Katrin Erk. 2016. PIC a different word: A simple model for lexical substitution in context. In Proceedings of NAACL-HLT. Association for Computational Linguistics.

A Kimball Romney, William H Batchelder, and Susan C Weller. 1987. Recent applications of cultural consensus theory. American Behavioral Scientist 31(2):163–177.

A Kimball Romney, Susan C Weller, and William H Batchelder. 1986. Culture as consensus: A theory of culture and informant accuracy. American Anthropology 88(2):313–338.

J Ruppenhofer, M Wiegand, and J Brandes. 2014. Comparing methods for deriving intensity scores for adjectives. In EACL 2014. Association for Computational Linguistics.

Edward Sapir. 1944. Grading: A study in semantics. Philosophy of Science 11:93–116.

Hans-Jürgen Sasse. 1993. Syntactic categories and subcategories. In Joachim Jacobs, editor, Syntax : ein internationales Handbuch zeitgenssischer Forschung (an international handbook of contemporary research), Walter de Gruyter, Handbcher zur Sprach- und Kommunikationswissenschaft (Handbooks of linguistics and communication science).

V Sheinman and T Tokunaga. 2009. AdjScales: visualizing differences between adjectives for language learners. IEICE Transactions on Information and Systems 92:1542–1550.

Vera Sheinman, Christiane Fellbaum, Isaac Julien, Peter Schulam, and Takenobu Tokunaga. 2013. Large, huge or gigantic? identifying and encoding intensity relations among adjectives in WordNet. Language Resources and Evaluation 47(3):797–816.

Chaitanya P Shivade, Marie-Catherine de Marneffe, Eric Fosler-Lussier, and Albert M Lai. 2015. Corpus-based discovery of semantic intensity scales. In HLT-NAACL. Association for Computational Linguistics, pages 483–493.

Il'ya Meerovich Sobol'. 1967. On the distribution of points in a cube and the approximate evaluation of integrals. Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki 7(4):784–802.

Tanya Stivers and Makoto Hayashi. 2010. Transformative answers: One way to resist a questions constraints. Language in Society 39(1):1–25.

U Sutrop. 1998. Basic temperature terms and subjective temperature scale. Lexicology 4:60–104.

Urmas Sutrop. 2001. List task and a cognitive salience index. Field methods 13(3):263–276.

O. Tange. 2011. Gnu parallel - the command-line power tool. ;login: The USENIX Magazine 36(1):42–47.

Margaret Thomas. 2011. Sībawayhi. In Fifty Key Thinks of Language and Linguistics, Routledge.

R L Trask. 1998. Key Concepts in Language and Linguistics. Routledge.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research 37:141–188.

Emiel van Miltenburg. 2015. Detecting and ordering adjectival scalemates. In MAPLEX.

Bob Van Tiel, Emiel Van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2016. Scalar diversity. Journal of Semantics 33(1):137–175.

Susan C Weller. 2007. Cultural consensus theory: Applications and frequently asked questions. Field methods 19(4):339–368.

Susan C Weller and A Kimball Romney. 1988. Systematic data collection. Sage.

Paul Westney. 1986. Notes on scales. Lingua 69(4):333–354.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. In ICLR.

Bryan Wilkinson and Tim Oates. 2016. A gold standard for scalar adjectives. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, pages 347–354.